A data analytic methodology for materials informatics

By

Osama Yousef Abuomar

A Dissertation
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
in Electrical and Computer Engineering
in the Department of Electrical and Computer Engineering

Mississippi State, Mississippi

May 2014

Copyright by

Osama Yousef Abuomar

2014

A data analytic methodology for materials informatics

By

Osama Yousef Abuomar

Approved:

_____

Roger L. King
(Director of Dissertation)

_____

Nicolas H. Younan
(Committee Member)

_____

Qian (Jenny) Du
(Committee Member)

_____

Hongjoo Rhee
(Committee Member)

_____

James E. Fowler
(Graduate Coordinator)

_____

Jason M. Keith
Interim Dean
Bagley College of Engineering

Name: Osama Yousef Abuomar

Date of Degree: May 16, 2014

Institution: Mississippi State University

Major Field: Electrical and Computer Engineering

Major Professor: Roger L. King

Title of Study: A data analytic methodology for materials informatics

Pages in Study: 158

Candidate for Degree of Doctor of Philosophy

A data analytic materials informatics methodology is proposed after applying different data mining techniques on some datasets of particular domain in order to discover and model certain patterns, trends and behavior related to that domain. In essence, it is proposed to develop an information mining tool for vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposites as a case study. Formulation and processing factors (VGCNF type, use of a dispersing agent, mixing method, and VGCNF weight fraction) and testing temperature were utilized as inputs and the storage modulus, loss modulus, and tan delta were selected as outputs or responses. The data mining and knowledge discovery algorithms and techniques included self-organizing maps (SOMs) and clustering techniques. SOMs demonstrated that temperature had the most significant effect on the output responses followed by VGCNF weight fraction. A clustering technique, i.e., fuzzy C-means (FCM) algorithm, was also applied to discover certain patterns in nanocomposite behavior after using principal component analysis (PCA) as a dimensionality reduction technique. Particularly, these techniques were able to separate the nanocomposite specimens into different clusters based on temperature and tan delta

features as well as to place the neat VE specimens in separate clusters. In addition, an artificial neural network (ANN) model was used to explore the VGCNF/VE dataset. The ANN was able to predict/model the VGCNF/VE responses with minimal mean square error (MSE) using the resubstitution and 3-folds cross validation (CV) techniques. Furthermore, the proposed methodology was employed to acquire new information and mechanical and physical patterns and trends about not only viscoelastic VGCNF/VE nanocomposites, but also about flexural and impact strengths properties for VGCNF/ VE nanocomposites. Formulation and processing factors (curing environment, use or absence of dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time) and testing temperature were utilized as inputs and the true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta were selected as outputs. This work highlights the significance and utility of data mining and knowledge discovery techniques in the context of materials informatics.

DEDICATION

I would like to dedicate this research to the love of my heart, my wife Walla, and to our cute precious baby girl, Salma.

ACKNOWLEDGEMENTS

The author wishes to express his warm thanks to Dr. Roger King for his dedicated and faithful help and continuous mentoring from the beginning to the end. Special thanks to Dr. Nicolas Younan for his valuable advices while pursuing this work. Sincere regards should be extended also to other committee members; Dr. Hongjoo Rhee and Dr. Qian (Jenny) Du for their cooperation and ultimate help during the organization phase of this work.

Also, sincere regards to Dr. Sasan Nouranian for his assistance in facilitating this research by providing the VGCNF/VE dataset required to conduct the study and for his assistance in reviewing and editing the research papers that have been written based on this work. Furthermore, the author wishes to express his acknowledgments to Dr. Jean-Luc Bouvard, Dr. Thomas E. Lacy Jr., and Dr. Charles U. Pittman, Jr. for their help and cooperation in reviewing and editing this work.

Finally, warm thanks to my mom, dad, brothers, and sister. I believe that without their continuous thoughts and spiritual help, this work would not have come true.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Background information

Informatics is a new area of interdisciplinary study that integrates information and computer sciences and a particular domain area to obtain comprehensive findings that assist in discovering new knowledge. Informatics is probably most synonymous with the biological sciences (bioinformatics), but it also has its role in materials. Materials informatics is a tool for material engineers who guide the whole process when different machine learning techniques are integrated with certain visualization human-like representation scenarios. In addition, this concept can minimize the handling of data as well as it makes the scientific research faster and easier.

This is fueled by the vast growth rate in information technology area and is considered the driving force of the knowledge discovery and representation applications, semantic technology, machine learning, data mining, information retrieval, etc., in the engineering disciplines.

For the field of materials informatics one of the driving forces will be spatio-temporal data mining and knowledge discovery. The importance of this approach can be recognized when one thinks of a sample that is undergoing some sort of stress over time. The surface being studied (2-D or 3-D) represents the spatial information of interest and as it is stressed this surface will change with time. Oftentimes, a physics based model

does not exist to explain or predict the observations. However, data does exist. It is with the utilization of knowledge representation schemes, sophisticated machine learning algorithms, insightful visualization approaches and materials science domain expertise that new knowledge may be discovered through the use of spatio-temporal data mining approaches on the observed data [1-3].

For this research, a materials informatics methodology is proposed after applying different data mining techniques on some datasets of particular domain in order to discover and model certain patterns, trends, and behavior related to that domain. In essence, it is proposed to develop an information mining tool for different materials ranging from polymers, like the vapor-grown carbon nanofiber/vinyl ester (VGCNF/VE) polymer nanocomposites analyzed and studied in this dissertation, to metal alloys [4-5].

1.2 Motivation

The initial attempt to utilize the concept of materials informatics and the underlying data mining and knowledge discovery techniques was to develop what is called "Third Generation (3G) Advanced High Strength Steels (AHSSs)" using datasets from conventional, first, and second generations AHSS.

However, due to the lack of samples of conventional, first and second generations of AHSS in literature and in experimental designs as well as the unclear view of the dataset dimensions[1] required to conduct the study, this attempt was unsuccessful to develop the 3G AHSS because there were not enough samples from which the desired properties of 3G AHSS can be inferred and utilized.

[1] The dimensions in data mining are the combination of both inputs and outputs of the developed model

Thus, the focus has been switched to a new class of advanced materials, nano-enhanced polymer composites and polymer nanocomposites [6], which have recently emerged among the more traditional structural materials such as steel and ceramics. Polymer nanocomposites have been used in a variety of light-weight high-performance automotive composite structural parts where improved specific properties and energy absorption characteristics are required [7]. Though polymer nanocomposites have recently been widely investigated [8, 9], they have never been studied in the context of materials informatics. Therefore, the purpose of this dissertation is to utilize data mining and knowledge discovery techniques with a thermosetting vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposite as the material system. Nouranian *et al.* [10-12] and Torres *et al.* [13] developed a relatively large and rich dataset for this material system, suitable for the proposed research. This dissertation seeks to use this dataset to demonstrate the usefulness of knowledge discovery and data mining techniques for nanocomposite material property characterization. In other words, to confirm that certain machine learning methodologies can provide similar outcomes as a trained domain specialist with this particular material system.

VGCNFs are commercially viable, nano-reinforcements with superb mechanical properties [14]. VEs are thermosetting resins suitable for automotive structural composites due to their superior properties in comparison with unsaturated polyesters [11-13, 15, 16]. Incorporating VGCNFs into VEs may provide improved mechanical properties relative to the neat matrix (i.e., specimens containing no VGCNFs). Data mining and knowledge discovery techniques can help discover and map patterns in the physical, mechanical, and system properties of VGCNF/VE nanocomposites, thereby

aiding the nanocomposite design, fabrication, and characterization without the need to conduct expensive and time-consuming experiments.

These mechanical properties, however, are dependent on the degree of VGCNF nanodispersion in the matrix achieved during the mixing stage of the process. Examples of good and poor nanofiber dispersion in the matrix are given in Figure 1.1, where two transmission electron micrographs of VGCNF/VE specimens are compared. Large nested groups of nanofibers (agglomerates) are a sign of poor VGCNF dispersion in the matrix, often resulting in inferior mechanical properties.

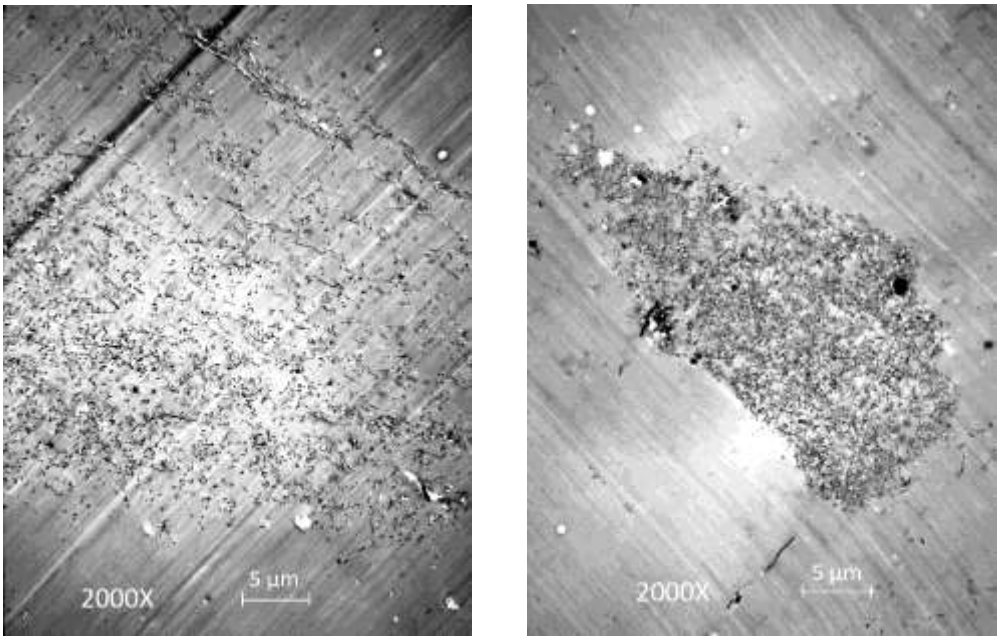


Figure 1.1 Transmission electron micrographs of two VGCNF/VE specimens

A nested VGCNF structure (agglomerate) is shown in the first image, indicating a poor VGCNF dispersion in the matrix, and a better-dispersed system is shown in the second image.

1.3 Research Contributions

The mindset of this research is to develop a data analytic materials informatics methodology by which different datasets related to a particular material system can be analyzed using a suite of data mining techniques and signal processing approaches in order to discover certain trends, patterns, and physical or chemical properties that were not known *a priori*. VGCNF/VE nanocomposites system is an example of such a material system and after reviewing published materials in the literature, most materials informatics applications involve only *metals* systems using supervised learning methodologies (see Chapter 2). However, polymer nanocomposites have never been studied in the context of materials informatics, so their patterns and their physical, mechanical, and system properties have never been explored using the empirical approaches, whether they are supervised and unsupervised learning approaches, utilized in the field of materials informatics. This substantiates the newness and genuineness of this dissertation whose main contributions will be:

- Developing a sensitivity analysis structure in order to discover the most and less dominant features, whether they are input design factors or output responses of the VGCNF/VE nanocomposites system.
- Developing a tool by which the VGCNF/VE specimens that yield the same storage and loss moduli responses can be extracted. This will facilitate computing and comparing the cost to fabricate these specimens by the domain experts[2].

[3] The cost data of VGCNF/VE specimens is not available but it's up to the domain experts to figure out the cost based on the results obtained from the developed tool.

- Developing a methodology by which the number of dimensions of a given dataset can be reduced and then by applying unsupervised learning approaches, different categories of VGCNF/VE specimens, like the neat VE specimens, along with their physical and mechanical properties can be categorized, identified, and validated based on the location of the results from the unsupervised learning.
- Developing a procedure where the output responses of VGCNF/VE system can be modeled based on the input design factors using a supervised learning approach to map input to output space.
- Establishing a baseline by which different datasets from multiple domains, related to different materials systems, can be combined (fused) in order to discover new trends and characteristics of the new materials system as well as to validate the facts mentioned in theory without the need to conduct expensive and time-consuming experiments.

Based on the above signal processing related approaches of the developed data analytic methodology of materials informatics, the following research papers have been published/ accepted:

- O. Abuomar, S. Nouranian, R. King (2014). **"Comprehensive Mechanical Characterization of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites Using Data Mining and Knowledge Discovery Techniques"**, *Journal of Data Mining and Knowledge Discovery*. (To be submitted).

- O. AbuOmar, S. Nouranian, R. King, J. L. Bouvard, H. Toghiani, T. Lacy, & C. Pittman (2013). **"Data Mining and Knowledge Discovery in Materials Science and Engineering: A Polymer Nanocomposites Case Study"**. *Advanced Engineering Informatics*. 27(4), 615–624. DOI:10.1016/j.aei.2013.08.002.
- O. Abuomar, S. Nouranian, R. King (2014). **"Classification of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites Using Support Vector Machines"**. (To be submitted).
- M. Al Boni, O.Abuomar, R.King (2014). **"ReShare: An Operational Ontology Framework for Research Modeling, Combining and Sharing"**. *The 2014 International Conference on Computational Science and Computational Intelligence (CSCI'14)*, March 10-13, 2014, Las Vegas, USA
- O. Abuomar, S. Nouranian, R. King (2013). **"Artificial Neural Networks Modeling of the Viscoelastic Properties of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites"**. *The 19th International Conference on Composite Materials (ICCM19), July 28-August 2, 2013*. Montreal, Quebec, Canada.
- O. Abuomar, R. King, H. Rhee (2012). **"Optimization of Artificial Intelligence Techniques for the Classification of 1st and 2nd Generation AHSS"**. *TMS Annual Meeting and Exhibition*, Stochastic Methods in Materials Research, March 2012, Orlando, FL

- R. L. King, O. Abuomar, H. Rhee, A. Konstantinidis, N. Pavlidou, M. Petrou (2011). **"On Materials Informatics and Pattern Formation in Materials"**. *ENOC 2011*, 24-29 July 2011, Rome, Italy.

1.4 Organization of Dissertation

The rest of the dissertation is organized as follows. Chapter 2 provides a literature review of previous data mining and knowledge discovery applications in the fields of materials science and mechanical engineering. Chapter 3 provides a technical description of the experiments and datasets of VGCNF/VE polymer nanocomposites system used in this research as well as it explains in details the theoretical aspects of the data mining and knowledge discovery techniques and algorithms implemented and utilized in this study. Chapter 4 shows detailed results after applying the respective algorithms and techniques based on the VGCNF/VE datasets being analyzed. In addition, Chapter 4 illustrates the results after analyzing AHSS data using the described data mining and knowledge discovery techniques and it also outlines the experimental requirements in order to design the 3G AHSSs. Chapter 5 shows how these results can be applied to the field of materials informatics. In other words, how these results make sense to the domain experts in order to validate their theories and experimental results or to discover new trends, patterns and behavior that are not known *a priori*. Chapter 6 provides an extended view on how different VGCNF/VE datasets can be combined into one unified nanocomposites framework. This has been accomplished by building a complete nanocomposites dataset that consists of viscoelastic VGCNF/VE dataset (studied in Chapters 3 and 4), VGCNF/VE impact strengths, and VGCNF/VE flexural dataset. This does not only test the validity and the effectiveness of the proposed data

analytics methodology, but also will allow materials scientists and other domain experts to more precisely characterize new mechanical and the physical patterns and trends of different nanocomposites systems such that they can determine the inputs' combination(s) that can yield a desired response(s). Chapter 7 concludes the dissertation and describes potential future work.

CHAPTER II

LITERATURE REVIEW

Materials informatics is a cross disciplinary effort combining materials science and information science [17]. It is a technique that aims at innovative materials development by way of informatics, which includes such elements as materials design using computational science and databases, a design of experiments methodology for populating databases, automatic combinatorial synthesis when searching for new materials, high throughput screening, compiling databases from obtained data, data sharing through networks, visualization of data, and data mining to predict different materials [18]. The overall goal of the data mining process is to extract information from a large complex dataset and transform it into an understandable structure, thus enabling knowledge discovery. This transformation of massive amounts of structured and unstructured data into information and then into new knowledge using a myriad of learning techniques is one of the great challenges facing the engineering community.

The following are some of the tools and resources; computational and experimental, that are required in order to understand all aspects of materials informatics [19]:

- The generation/creation of data which includes the following techniques:

    - Combinatorial experiments which has the advantage of high throughput.
    - Systematic data collection; this technique is very slow.
    - Dynamic experiments, which is ideal for time series data.
- Data warehousing or data correlations.
- Dimensionality reduction. This technique offers:
    - Identification of the most dominant trends and patterns in the data.
    - Data variability can be easily analyzed and visualized by smaller number of dimensions.
    - Noise and outliers reduction such that implementing data mining and data analytics techniques and algorithms will be feasible and more beneficial.
- Data mining, which includes the following steps:
    - Unsupervised and supervised learning methodologies.
    - Predictive modeling techniques. These techniques look for hidden data trends and process-structure-property relationships.
- Knowledge discovery.
- Visualization techniques.
- Cyber infrastructure.

The materials informatics cycle is shown in Figure 2.1 [19]:

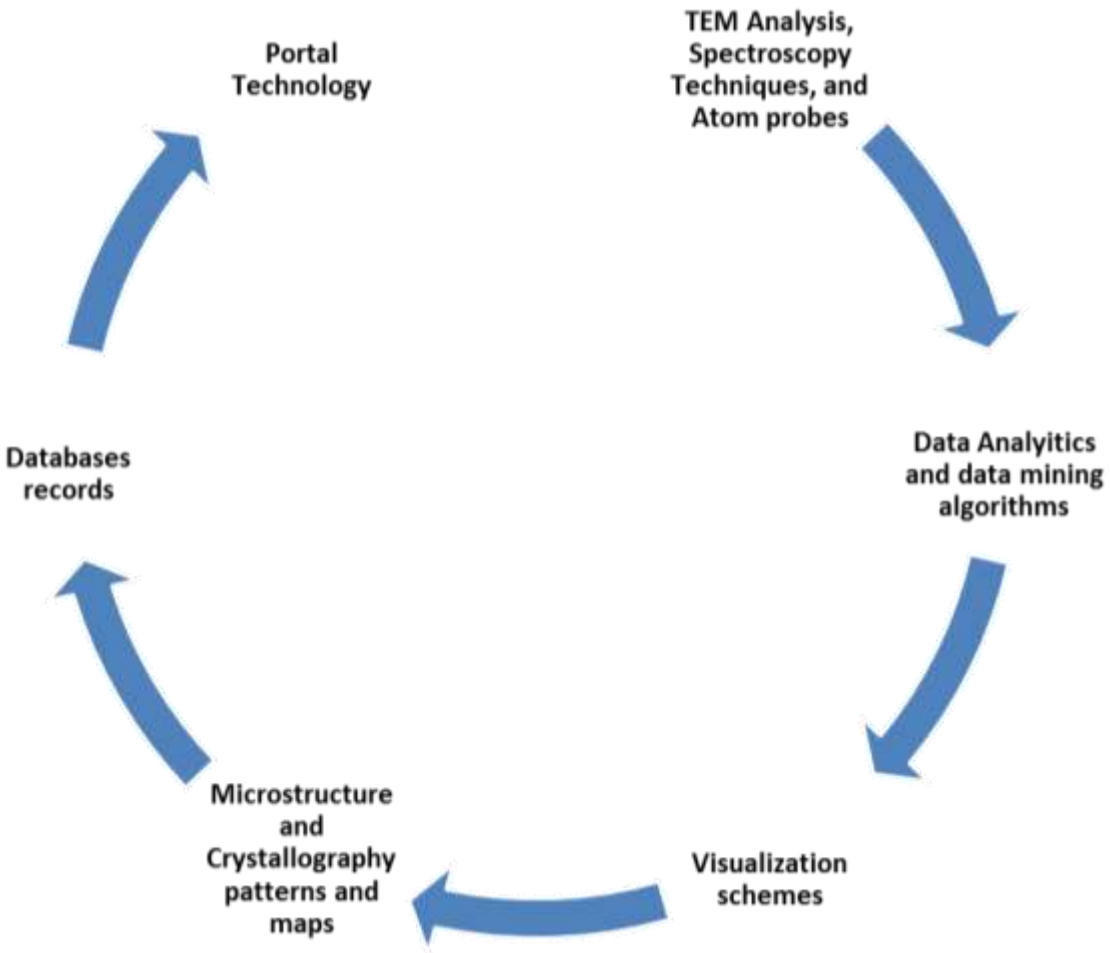


Figure 2.1 The main steps in materials informatics cycle [19].

In combination with data dimensionality reduction, implementing higher-order statistical tools on missing, skewed, and ultra large scale datasets are effective. They can also develop some models for uncertainty like fuzzy clustering and clustering analysis, partial least squares methods, support vector machines (SVMs) and singular value decomposition technique [19].

There are several applications and approaches in the literature utilized for materials informatics. Hu, C *et al.* [20] have used materials informatics to resolve the

problem of materials science image data sharing. They have presented an ontology-based approach that can be used to develop annotation for non-structured materials science data with the usage of semantic web technologies. In their work, they have stated that the role of ontology-based techniques is to exploit a specialization representation of materials science image content, and to form a knowledge base of non-structured materials data. Their approach is implemented by a four-layer architecture, which includes ontology building, semantic annotation, reasoning service, and application. As an example, they took metallographic image data and built a metallographic image using Ontology Web Language (OWL-ontology).

Paolini, C. and Bhattacharjee, S. suggested a purely technical platform for computing the thermochemical properties of substances [21]. By using specific temperature, database primary key, phase, species name, and Chemical Abstract Service (CAS) number [22]. They proposed a thermochemical Web Service that has different methods to calculate and return different thermodynamic characteristics. This Web Service can be invoked from Java and JavaScript using an Asynchronous JavaScript and XML (AJAX) technique. All of this computational work can be implemented in a common applications such as Microsoft Excel, Matlab, Mathematica, R, …etc. Using the developed Web Service, A joint Army Navy Air Force (JANAF) table, commonly used by scientists and engineers, can be easily generated [21].

Sabin, T. J. *et al.* [23] utilized Gaussian process framework as a statistical technique to predict the outputs based on a probability distribution function over the training dataset. This framework can be trained easily based on the available input variables, and then it can be easily tested in order to make the corresponding predictions

and estimations. The dataset used for training and testing consists of samples of high-purity Al alloy with 1 wt. % Mg (annealed Al-Mg and cold alloy).

Their work predicted the microstructure evolution arising from static recrystallization using non-uniform straining conditions. Thus, strain, temperature (^{o}C), and annealing time (s) were the inputs of the model and the mean logarithm of grain size (D) was its output.

According to Sabin, T. J. *et al.* [23], the Gaussian framework predicted the Gaussian distribution's mean as well as the Gaussian's standard deviation. These predictions are considered quantitative measures by which calculating a probability distribution fit is possible. This study used the *likelihood* measure which a data prediction strength statistical measure. That is, when estimating the model's performance, instead of just taking the difference between the predicted and target values, *likelihood* measure calculates the predictions' probabilities. The likelihood log (L) is:

$$L = -\sum_{k}\left[\frac{(\tau_k - p_k)^2}{2e^2{}_k} + \ln e_k\right] \tag{2.1}$$

where k and e_k are the dataset's vectors (points) and the Gaussian's standard deviation at the k^{th} vector. τ_k and p_k are the target and predicted values at the k^{th} point respectively.

The advantage of the Gaussian process framework is that it can evaluate a joint probability distribution function for all output responses using all input variables. However, other techniques like artificial neural networks (ANNs) use empirical mappings in order to predict one output value for a given set of input variables (parameters).

Roberts, K. *et al.* [24] presented a model that classifies different materials based on their microstructure. This was performed in a mterialographical laboratory designed for such tasks. The core of the designed model is a support vector machines (SVMs) classifier that identifies the appropriate class of given material sample. The classification is performed using microstructural characteristics and properties from particular image processing. These features include Haralick variables, and stereological features such as Euler parameter and fractal dimension.

According to Roberts, K. *et al.* [24], the SVMs classifier uses numerical feature vectors. There are 20 different features (*m*) that characterize pixels and their statistical properties in a microstructural map. The SVMs classifier uses 14 Haralick variables that extract the textural properties of the map, and 6 stereological variables that extract the stochastic geometry of the microstructural images.

Furthermore, the system supported two kinds of sensitivity analyses [24]. The aspects of these analyses includes:

- Starting from a small training dataset where its size is increased with time. The precision, recall rate, and accuracy percentage are calculated at each step.
- All features are used in training at the beginning and the resulting classification rate is calculated. At each step, some features are removed from the analysis and the corresponding classification rate is calculated. The second aspect determines which features are the most important for specific classification tasks.

The attempt to design an ontology for an agent-based system of the Materials Microcharacterization Collaboratory (MMC) was explained [25]. The materials ontology expressed some rules and directions for technicians in order to use some scientific instrumentation such as, electron microscopes and a neutron-beam line device.

In order to represent the components of the task performed by the scientists and users, the domain ontology built some units. This was done in the access phase of the project. In addition, the project contained knowledge related to the scientists and domain experts as well as the instrumental variables used in the project.

Four main entities or units in the domain ontology were created: User, Expert, Instrument, and Experiment [25]. The User unit specifies the experiment's details that have to be conducted, the instrument's properties and a unique user identifier (UID) which was previously issued. The Instrument unit specifies the Expert who owns management and control over the instrument, instrument properties, access methods, the type of training required in order to use the instrument, availability, and location [25]. The Experiment unit specifies some properties related to the projected experiment, some inner properties of the sample material to be examined, some properties (e.g. temperature, density, etc.) related to the environment in which the experiment is to be conducted, the required experiment's instrumentation, and the influence of the projected experiment on the area/ field that has been studied and analyzed.

On the other hand, Belov G.V. and Iorish V.S. [26] presented a technical proposal for the data exchange among databases on thermodynamic properties of substances. They stated that the data exchange information should contain:

- Data model: in the model, structure and composition of the data (including field names, types, description, units, and constraints) should be defined.
- Data file(s) description using extensible mark-up language (XML) or any other language.
- Data file(s) itself.

Furthermore, the following ways to simplify exchanging the materials' thermodynamical data were suggested [26]:

- Developing the principles of data field name forming (naming conventions).
- Developing the list of names for all data stored in databases and make this list available via internet.
- Developing agreement on physical data units.

Guessasma and Coddet [27] dealt with 13 wt.% alumina and titania microstructural characteristics that were coated under different APS[3] scenarios. ANNs were utilized using the atmospheric plasma spray parameters. These parameters are directly linked to alumina, titania, un-molten particle concentrations, and porosity. Thus, the developed ANN model was used to analytically measure the results after providing the process parameters to the feature structures.

Particularly, the input parameters used in the ANN model were: the current used in arc (I), flow rate of plasma gas ($H+A$) and hydrogen ratio (H/A) [27]. Each of these parameters was applied as an individual neuron in the model. The output responses were

[3] APS conditions refer to the conditions and standards set by American Physical Society.

un-molten particle ratio (*ln*), porosity level (*P*), titania (*T*) and alumina (*A*) phase concentrations.

The architecture of the ANN used in the study as well as a flowchart of how the ANN was trained and optimized in the study are shown in Figure 2.2 [27].

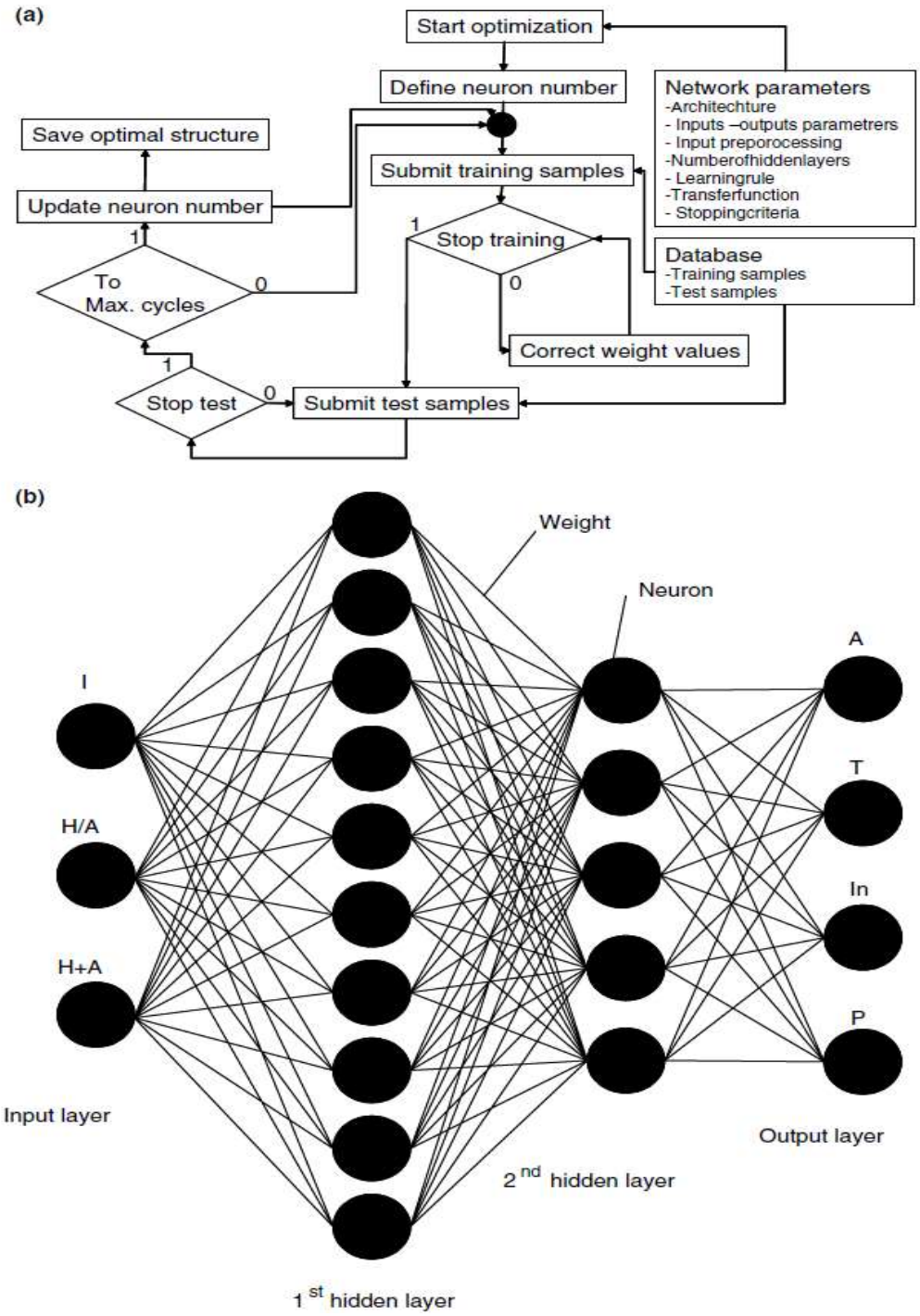


Figure 2.2 (a) Optimization process diagram. (b) An ANN model where microstructural relationships were predicted.

The APS variables were used as the model's inputs.

Swaddiwudhipong, S. *et al.* [28] utilized another important and efficient materials informatics technique. This technique called least squares, support vector machines (LS-SVM). Four LS-SVM models were designed in order to create a relationship between the indentation load-displacement characteristics and the properties of elasto-plastic material which are subject to the law of power hardening without using any iterative resorting approaches and it was discovered that LS-SVM was robust in determining the parameters in this relationship. Using load-displacement curves which were not used in learning and testing, the sample material's properties were successfully predicted by the final adjusted LS-SVM framework given that new sets of these curves were applied to the framework.

Analytically, solutions of closed relations (functions) are not trivial to be established because the resulting nonlinear optimization task is very difficult to solve [28]. Another option is to use functional approximation procedures in order to establish the relations between the sample material's features and the load parameters and then tune them numerically. This can be easily handled through using an ANN architecture. A single indenter's findings, Huber *et al.* [29, 30] showed that ANN can be used to characterize the material of thin film samples on a substrate.

According to Swaddiwudhipong, S. *et al.* [28], local minima problem was exhibited in the traditional ANN architectures and their training methods. In addition, to avoid the problems of over-fitting and under-fitting, the number of neurons used in the model has to be carefully adjusted. In this study, LS-SVMs models were utilized to characterize the material's features based on various geometric dual indenters' structures using load-displacement response [31]. Finite element analyses (FEA) were used to generate the adopted data and to characterize the response of elasto-plastic materials

which are subject to the hardening-strain law during the process of indentation of various geometric structures.

SV regression establishes the relations:

$$f(x) = \langle x, w \rangle + c \tag{2.2}$$

The constrained optimization problem is to minimize

$$\frac{1}{2}\|w\|^2 + k\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2.3}$$

Subject to

$$y_i - \langle xi, w \rangle - c \leq \varepsilon + \xi_i \tag{2.4}$$

$$\frac{1}{2}\|w\|^2 + k\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{2.5}$$

$$\xi_i, \xi_i^* \geq 0 \tag{2.6}$$

The constant *k,* which must be greater than 0, determines the trade-off between the function *f* flatness and the tolerated amount of deviations larger than ε [28]. The slack variables, ξ_i and ξ_i^* can be used in order to conduct the convex optimization analysis. By applying a linear regression algorithm, the optimization task can be mapped into a higher dimensional feature space. This can be done by using the appropriate kernel functions in the analysis. Thus, nonlinear function approximation is resulted in the input space [28].

Pelckmans *et al.* [32] implemented LS-SVMs and made the package available for scientific research purposes. The URL of the package is:

www.esat.kuleuven.ac.be/sista/lssvmlab

The implemented LS-SVMs [28] were designed using LS-SVM, lab1.5 Matlab toolbox, version 6.5 [32]. To design an efficient LS-SVM model, satisfactory values of the two variables γ and σ^2 have to be computed. γ is a control regularization variable that is related to the parameter *k* in Equation 2.3 and it determines the minimization and smoothness behaviors of the fitting error. The control variable σ^2 is the kernel's bandwidth. This problem is a minimization task of the cost function *f*. The *tunelssvm* function integrates the optimization analysis of grid-search to search for the optimal values of γ and σ^2 in the problem's domain.

Mukherjee, M *et al.* [33] developed an ANN to predict the retained austenite's strain induced transformation characteristics. There were 13 inputs parameters of the model including matrix microstructural features and forming properties, steel's chemical structure, and initial austenite concentration.

The ANN predictions in [33] verified the importance of carbon on the stability of retained austenite. In addition, it showed the insignificant effect of manganese. Although the training and test datasets were insufficient, the developed ANN model was able to show the complex effects of microstructural features and temperature. In addition, as a new contribution, the retained austenite stability was analyzed using matrix microstructural features which were accurately quantified in order to be analytically used in the study.

As a future work, the sensitivity analyses of the model can be extended using a larger and richer database. That is, important features like morphology and retained austenite size, austenite carbon concentration, and the stress status of the deformation

tests. However, this database can be generated by utilizing suitable reporting services and proper experimentation procedures [33].

The dataset of Charpy toughness for submerged arc weld metal and manual metal arc were analyzed using an ANN architecture using Bayesian approach [34]. In the developed model, the Charpy toughness can be viewed as a general empirical representation of parameters that are vital in characterizing and designing the steel welds. These variables are listed in Table 2.1.

Table 2.1 The implemented input parameters.

Variable	Range	Mean	Standard Deviation
Process	Submerged Arc Manual Metal Arc		
Yield Strength MPa	347–645	471	12.7
Carbon wt.%	0.029–0.13	0.08	0.004
Silicon wt.%	0.28–1.14	0.49	0.05
Manganese wt.%	0.77–2.50	1.32	0.07
Phosphorus wt.%	0.008–0.028	0.015	0.001
Sulphur wt.%	0.002–0.017	0.010	0.0005
Aluminium wt.%	0.001–0.04	0.014	0.002
Nitrogen p.p.m.w.	26–119	67	4
Oxygen p.p.m.w.	234–821	412	30
Primary Microstructure %	0–91	34	4
Secondary Microstructure %	9–100	66	2
Allotriomorphic Ferrite %	16–62	31	2
Acicular Ferrite %	11–81	55	2
Widmanstätten Ferrite %	0–35	14	2
Temperature K	213–293	259	25
Charpy Toughness J	4–215		

Note that the chemical composition was measured in wt. % or in parts per million weight (p.p.m.w.) [34].

Yescar, M.A. and Bhadeshia, H.K.D.H. [35] designed a model that predicted the maximum retained austenite volume fraction in ductile irons structures. The model's input parameters are the chemical structure and the heating conditions.

According to them, there are two consequences that determine the model's behavior. The first consequence was that more austenite structures were stayed at ambient and cooling temperatures because bainite structures were formed and as a result this enriched the residual austenite. The second consequence can be noted at low temperatures because at the beginning of the transformation, bainite formation resulted in lower amounts of austenite that can be used for retention.

The work done by Yoshitake, S. *et al.* [36] was to use an ANN in order to model the changes of austenite γ and retained austenite γ' lattice parameter stages using the chemical composition and test temperature as input parameters. Reasonable predictions were made for different alloys that match X-rays measurements. The variation of the lattice constant which is directly proportional and dependent on the individual alloying elements contents and on test temperature can now be analyzed using other computer programs which deal with the partitioning of solutes between the γ and γ' phases.

The architecture of the ANN model used is shown in Figure 2.3.

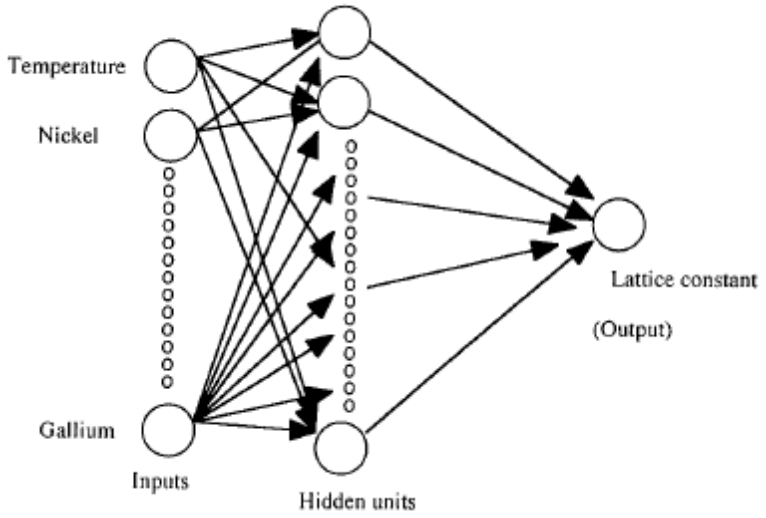


Figure 2.3 The ANN structure of the study [36].

Using chemical structure and processing parameters, Tancret, F *et al.* [37] utilized Gaussian processes to analyze the properties of tensile and creep rupture of alloys by using of large datasets on different alloys. Using different values of stress and temperature, the developed models predicted the actual effect of alloying elements on some of the mechanical properties of alloys.

An ANN models was used to analyze the conditions that affect weld mechanical attributes of pure titanium [38]. The models predicted the yield strength, ultimate tensile strength, area reduction, Vickers hardness and Rockwell B hardness, and elongation. Input variables were generated from mechanical testing of single-pass welds. It was confirmed that nitrogen and oxygen are the most important variables for mechanical properties, and the hydrogen has the least influence while the cooling rate is more significant than carbon and iron in the UTS model, and more significant than oxygen and iron, and equally significant with carbon in the yield strength model.

An ANN model was developed to predict the strength where 108 variables including different rolling variables and chemical structure of steel were used as the model's input variables [39]. This work was restricted to steels with a ferrite and pearlite microstructure. By varying carbon and manganese contents, the model was able to estimate different results of the yield strength and tensile strength ratios.

A Bayesian ANN model was trained. That is, ultimate tensile strength, yield strength, and elongation percentage were analyzed using the model, so tensile property data for mechanically alloyed, oxide dispersion strengthened ferritic stainless steels were thoroughly examined [40].

Narayan, V *et al.* [41] used a feed forward ANN to model the dependence of the stress on strain. The model's input variables include the chemical structure and the testing temperature.

The stress-strain graphs were shown to be [41]:

- highly dependent on temperature.
- less sensitive to the interpass time with increasing temperature.
- not very sensitive to chemical composition at very high temperatures.

Ryu, Joo, and Bhadeshia, H.K.D.H. [42] dealt with big hot-rolled steels and designed an ANN architecture to predict the strength of hot-rolled steels that have a mixture of ferrite and pearlite. The inputs were chemical structure and process parameters. The architecture was then matched to the corresponding datasets in order to highlight the importance of carbide formers addition in altering the attributes and the characteristics of steels.

Gavard, L *et al.* [43] used ANN analysis to investigate the austenite's formation during the steels heating. A comprehensive database that includes the model's parameters of chemical structure, the heating rate, and the Ac_1 and Ac_3 temperatures was utilized. The ANN's findings matched the results from the theory of phase transformation technique.

According to [43], the temperature was the ANN's response and the chemical structure of steel and the heating rate were the model's input parameters. The modeled temperature is the starting point of the first formation of austenite during heating and the austenite's transformation. The model predicted the transformation temperatures fairly well (about 95% confidence interval, or +/- 40K temperature range). This proved that using the Ac1 temperature resulted in less reliable predictions than that of the Ac_3 temperature. This due to the fact that the Ac_1 temperature was most likely more difficult to be experimentally measured or there is a non-equilibrium phase transformation.

The Charpy and tensile characteristics for irradiated low activation martensitic (LAM) steels using ANN approach were modeled [44]. The study focused on Eurofer (9-Cr) and F82H (8-Cr) as they are the most common LAM steels and have a variety of potential future applications. The analysis of ANN's outputs showed that ductile-brittle transition temperature (DBTT) and YS are sensitive to all inputs. This significantly high-lighted the importance to record the complete set of experimental data.

Bayesian ANN architecture was developed to model the growth rate of fatigue using nickel base super alloys [45]. The model had 51 input variables, including chemical structure, applied frequency, temperature, grain size, applied waveform load, stress intensity range (Δ K), log Δ K, heating conditions, atmosphere, R-ratio, long and short

cracks growth distinction, and strength and measured thickness. A sensitivity analysis was conducted in order to improve the predictions' accuracy. The architecture of the ANN used in the model was shown in Figure 2.4 [45].

A sensitivity analysis was conducted based on the model [45]. As an agreement with Paris law, it was verified that $\log \Delta K$ is more related to the fatigue crack growth rate than to stress intensity range. In addition, when taking the grain size as a separate input, one can determine its effects on the model. Particularly, it was verified that the fatigue crack growth rate was decreased when the grain size was increased. This demonstrated the ability of this method to reveal new phenomena cases where experiments cannot be designed to study each variable in isolation.

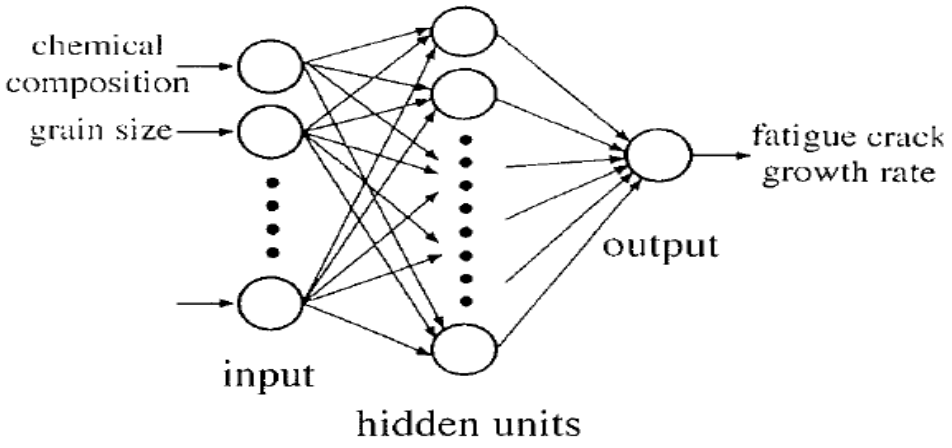


Figure 2.4 ANN model used in the study [45].

Forsik, S. and Bhadeshia, H. [46] model the elongation of neutron-irradiated steels using an ANN structure. The ANN Predictions matched the experimental values. The ANN architecture was extended to predict the elongation response at high doses and

high temperatures (200 dpa irradiation and 750°C respectively). Predictions made by the model contain very large uncertainties due to the lack of experimental samples at some doses and temperatures.

The ANN model was trained on an output Y :

$$Y = \ln(-\ln\left(\frac{y - y_{\min}}{y_{\max} - y_{\min}}\right)) \tag{2.7}$$

where y_{max} and y_{min} are the maximum and minimum responses in the database respectively (i.e., the model was trained using a boundary function of the output).

According to [46], the better mechanisms used for hardening and the availability of more irradiation experiments at high doses and temperatures will help to limit modeling uncertainties and increase the ANN model's reliability when used for unseen data samples. In addition, models are typically trained on the output directly rather than on a function of the output.

Cottrell, G.A *et al.* [47] designed a Bayesian ANN system to estimate the neutron irradiation change of the Charpy temperature during the transformation of ductile to brittle (*ΔDBTT*). This was accomplished using 40-D dataset describing low activation martensitic steels. The irradiation concentrations were less than 100 displacements per atom (dpa). The conducted sensitivity analysis demonstrated the importance of irradiation temperature in calculating *ΔDBTT*.

The developed model confirmed some well-established and known relationships as well as it discovered some hidden patterns, trends and characteristics [47]. The following are some of them:

- When sensitivity analysis was conducted, the input variables that significantly control *ΔDBTT* were determined. The most important input is T_{irr} followed by the function $(dpa)^{1/2}$. In addition, chemical structure (Cr, C and other components) and heat treating conditions in before the irradiation phase are important.
- While T_{irr} was increased above about 450°C, *ΔDBTT* was decreased and saturated.
- Material Samples with high doses and high value of T_{irr} are easy to recover.
- After irradiation, the lowest value of *ΔDBTT* was obtained with Cr contents.
- After irradiation, it was proved that *ΔDBTT* was decreased with Ta contents.

Charpy toughness is a quality control parameter that is important in fractures as it is an indicator of the absorbed energy by the material sample [48]. If the absorbed energy is high, then the sample will be less vulnerable to brittle fracture. Thus, charpy toughness can be used to determine the integrity of components in the sample [48].

According to [48], this energy is always positive as it is impossible for the sample to emit energy to the testing machine. An empirical model to calculate the Charpy impact toughness of steel welds as a function of composition and processing was used [48]. Also, this work discovered patterns in the weld metals Charpy toughness energy when some solutes and welding variables were used as the model's input parameters.

The dataset used in creating the ANN models used in the study is illustrated in Table 2.2 [48].

Table 2.2 The dataset used in creating the ANN model [48].

Variable	Range	Mean	Standard deviation
C/wt%	0.008-0.19	0.07	0.02
Si/wt%	0-1.63	0.35	0.14
Mn/wt%	0-2.31	1.20	0.40
S/wt%	0.002-0.14	0.01	0.01
P/wt%	0-0.25	0.01	0.01
Ni/wt%	0-12.40	0.88	1.83
Cr/wt%	0-19.50	0.35	1.19
Mo/wt%	0-2.43	0.20	0.31
V/wt%	0-0.53	0.01	0.03
Cu/wt%	0-2.18	0.08	0.21
Co/wt%	0-0.092	0.003	0.01
W/wt%	0-3.86	0.004	0.11
O/ppmw	25-1700	429	161
Ti/ppmw	0-770	71	115
N/ppmw	0-1000	95	67
B/ppmw	0-200	8	27
Nb/ppmw	0-1770	32	109
HI/kJmm^{-1}	0.21-16.36	1.49	0.83
IT/℃	20-350	182	39
PWHTT/℃	20-940	198	287
PWHTt/h	0-100	1.3	4.2
D_{Fe}	0-3.68 × 10^{12}	6.22 × 10^{10}	4.47 × 10^{11}
TT/℃	-196-136	-34.5342	36
Charpy toughness/J	0.1-356	85	50

ppmw: Part per million by weight
HI: Heat input IT: Interpass temperature
PWHTT: Post-weld heat treatment temperature
PWHTt: Post-weld heat treatment time
D_{Fe}: A variable for iron diffusion during post-weld heat treatment
TT: Test temperature

Sourmail, T *et al.* [49] applied an ANN model in order to predict the life and the stress of creep rupture. The ANN's inputs were the chemical structure, testing conditions,

temperature of solution treatment, and the stabilization ratio. The model was applied to a comprehensive database with different austenitic stainless steels compositions, structures, and heat treatments.

In the developed model, a sensitivity analysis was conducted such that the importance and the interactions between different inputs were clearly identified and thoroughly analyzed.

TRIP steel was fabricated using ANN and genetic algorithms. In the design, the silicon percentage was low [50]. The study shows a new finding after using the combination of ANN and genetic algorithms.

In TRIP assisted steels, if the concentration of silicon is ~1.5 wt%, then it will limit the ability of cementite to precipitate during the development phase of of bainitic ferrite. As a result, the carbon remains into the residual austenite which will make the austenite to be stable around the ambient temperature [50].

The study's goal was to design and verify a framework which allows the domain experts to determine the alloying elements' combination that can achieve not only the optimum quantity of retained austenite and the silicon concentration in the resulting structure [50]. Furthermore, this study mentioned that the dataset used to design the ANN model were taken from the literature. The variables used are listed in Table 2.3 [50]. The concentrations of chemical elements are expressed in wt-%.

Table 2.3 Data used for modeling retained austenite fraction.

Parameter	Min.	Max.	Avg.	σ
Carbon	0.0950	0.3920	0.2102	0.1020
Manganese	0.6000	1.9900	1.4057	0.2796
Silicon	0.0400	2.1000	0.9619	0.5369
Aluminium	0.0000	2.0000	0.2979	0.5236
Phosphorus	0.0000	0.2040	0.0223	0.0401
Molybdenum	0.0000	0.1400	0.0043	0.0218
Copper	0.0000	0.5100	0.0065	0.0573
IA temperature	730 ºC	840 ºC	778.2653 ºC	25.7829
IA time	60s	450s	213.2143s	110.8414
IT temperature	300 ºC	525 ºC	404.2347 ºC	39.5554
IT time	0s	5400s	441.3673s	588.7228
Retained austenite volume content (%)	0.12	27.07	10.3805	5.8054

IA stands for intercritical annealing, IT for isothermal transformation to bainite and σ is for the standard deviation of the datset [50].

Good tensile properties were achieved when the alloy was transformed into a structure that consists of a combination of bainitic ferrite, δ ferrite dendrites, and carbon retained austenite [50]. At ambient temperature, the elongation was uniform at the level of 23%.

On the other hand, Gaussian analyses were utilized to estimate the polycrystalline physical properties using nickel based alloys where the chemical structure and the heating conditions were the model's inputs [51]. After applying the model, the respective metallurgical trends were reproduced and the behavior of the new developed alloys was generated, so the study developed a framework where creep stress of nickel based alloys and yield strength were predicted for power plant applications where the cost was reduced significantly [51].

According to the study, new alloys were designed using the Gaussian modeling method which is used to present complex and multidimensional trends in experimental dataset in an effective manner.

Das *et al.* [52] developed an ANN model which included the steel's chemical structure, the strain before aging as well as aging conditions and annealing behavior. The developed model was used to investigate some anomalies in literature. That is, although the deformation and aging behaviors can be rationalized, further experimental future work can be accomplished because niobium genetic structure was associated with sound uncertainties. With the ANN model it was possible to rationalize the effects of the "pre-strain" and aging temperature on the extent of bake hardening.

In the work presented by Brun, F *et al.* [53], developed and ANN architecture to predict creep rupture strength of martensitic and bainitic steels with Fe-2.25Cr-1Mo and Fe-(9-12)Cr structure. The ANN's inputs were chemical structure, heat processing and time at particular temperature. This architecture along with different thermodynamic and kinetic calculations, was utilized to present two new alloys of 10CrMoW which have superior creep rupture characteristics. The parameter selected to be modeled was the rupture stress not only because of the availability of the corresponding datasets in literature, but also because it is a vital variable used in industry during alloy manufacturing processes [53].

Metzbower, E.A. *et al.* [54] conducted a sensitivity analysis using an ANN model in order to determine if yield strength or ultimate strength plays a significant role in predicting the ductility and fracture toughness if used with chemical structure and cooling rate as other input parameters of the ANN model.

According to Metzbower, E.A. *et al* [54], after the sensitivity analysis procedure, chemical structure and cooling rate were found to be the most important features in modeling the mechanical properties of a weld metal. In addition, yield strength or ultimate strength are not significant in the ANN model and their presence as inputs didn't provide any trends or structure that relate them to the whole model. However, some information about the yield or ultimate strength is needed before conducting the study.

An ANN model was developed and showed that if Ni and Mn concentrations were controlled with respect to each other, the high-strength steel weld metals strength could be enhanced. However, yield strength would be moderately affected by such scenario. That is, the developed model's task was to increase toughness when low temperatures were applied [55].

The trend of predicted weld toughness at - 40 °C where Mn and Ni contents were used as inputs to the ANN model was shown in Figure 2.5 as contours. By reducing Mn contents for Ni concentrations above 5 wt. %, the predicted toughness by the model could be increased significantly while keeping yield strength at an acceptable level [55].

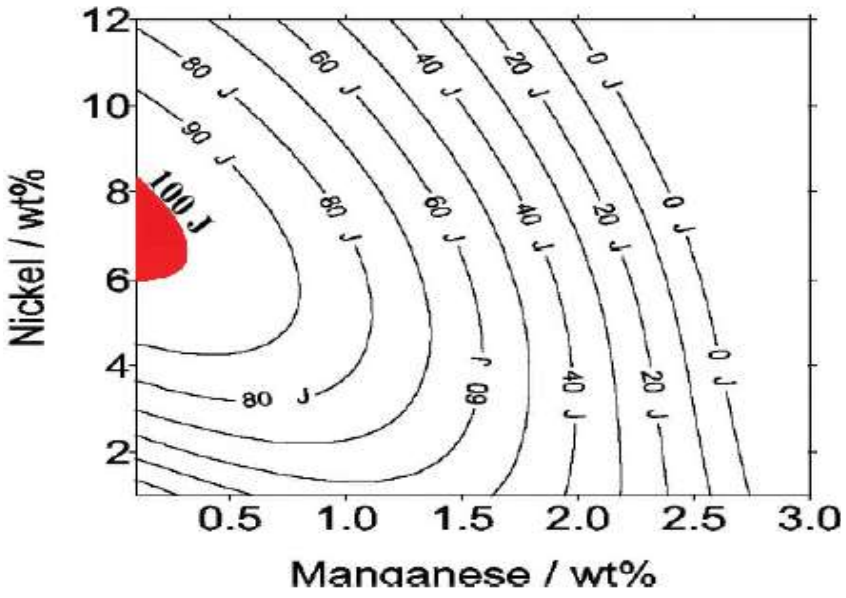


Figure 2.5 The trend of predicted weld toughness at - 40 °C

Mn and Ni contents were used as inputs to the ANN model.

Dimitriu, R. C. and Bhadeshia, H. K. D. H. [56] developed an ANN model to predict the 0.2% strength of creep steels of resistant ferritic microstructure where chemical structure and heating parameters were used as inputs. After conducting a sensitivity analysis and by combining the developed model with other findings from literature, it was shown that there was a point within 780–840 K range beyond which the yield strength didn't have any significant importance in the resulting microstructure.

The variables used to develop the ANN model were shown in Table 2.4 [56].

Table 2.4 Variables used to develop the ANN model [56].

Variable	Minimum	Maximum
Aluminium, wt-%	0·001	0·04
Carbon, wt-%	0·09	0·48
Copper, wt-%	0·0001	0·25
Chromium, wt-%	0·0001	12·38
Manganese, wt-%	0·38	1·44
Molybdenum, wt-%	0·01	1·05
Nickel, wt-%	0·0001	0·6
Nitrogen, wt-%	0·001	0·04
Silicon, wt-%	0·18	0·86
Austenitising time, min	10	5400
Tempering time, min	30	660
Austenitising temperature, K	1143·15	1243·15
Tempering temperature, K	898·15	1023·15
Test temperature, K	293·15	973·15
Hot strength, MPa	69	660

A large complex strength steels dataset was presented and analyzed [57]. The inputs parameters were the chemical structure, heating conditions and testing temperature. The study compares two techniques; a Bayesian ANN framework and genetic algorithms in which the data are organized in an evolutionary manner. It was concluded that ANN was able to analyze complex data more thoroughly whereas genetic algorithms were not as accurate and sophisticated because they require further learning and intervention to obtain acceptable results.

The variables used to develop the models were shown in Table 2.5 [57].

Table 2.5 Variables used to develop the models [57].

Variable	Minimum	Maximum
Aluminium / wt%	0.001	0.04
Carbon / wt%	0.09	0.48
Copper / wt%	0.0001	0.25
Chromium / wt%	0.0001	12.38
Manganese / wt%	0.38	1.44
Molybdenum / wt%	0.01	1.05
Nickel / wt%	0.0001	0.6
Nitrogen / wt%	0.001	0.04
Silicon / wt%	0.18	0.86
Austenitising time / min	10	5400
Tempering time / min	30	660
Austenitising temperature / K	1143.15	1243.15
Tempering temperature / K	898.15	1023.15
Test temperature / K	293.15	973.15
Hot strength / MPa	69	660

Javadi and Rezania [58] provided a coherent architecture based on evolutionary polynomial regression-based constitutive model (EPRCM) for complex materials modeling. Finite element analysis (FEA) was utilized in the framework. Thus, the EPR relationships were merged into the FEA model. This resulted in an intelligent technique that can be used in the developed architecture.

According to the study, instead of using the traditional constitutive models, the EPRCM model can be used for different materials. After comparing the results and analyses of the developed architecture with those of conventional FEA analyses, EPRCMs were able, with high accuracy, to identify the behavior of different materials and hence they can be applied in FEA models.

Using the patterns (clusters) in each material's image, Brilakis *et al.* [59] presented an automated and content-based construction site to retrieve different images. In this method, separate clusters were identified by grouping each image pixels together and then were compared with different material samples. Therefore, by using the contents of each image, domain experts were able to search for different construction images. Sharif Ullah and Harib [60] proposed a procedure in order to solve the problems of selecting different materials where the design specifications and information and working circumstances are not previously known. The procedure input parameters were; a verbal explanation of the problems by describing the importance of materials attributes and properties, and the corresponding graphs of the verbal explanation of the problem. Using this method, domain experts can select the best materials to design different robotic links. Specifically, it was discovered that composite materials are the best option to design such links.

CHAPTER III

METHODS, ALGORITHMS, AND TECHNIQUES

In this dissertation, several signal processing and supervised and unsupervised knowledge discovery techniques were used to explore a large VGCNF/VE dataset [10]. The dataset consisted of 240 data points each corresponding to the combinations of five input design factors and three output responses, i.e., a total of eight "dimensions." The dimensions in data mining are the combination of both inputs and outputs of the developed model. The dimensions of the VGCNF/VE dataset are VGCNF type, use or absence of dispersing agent, mixing method, VGCNF weight fraction, temperature, storage modulus, loss modulus, and tan delta (ratio of loss to storage modulus), where the last three dimensions correspond to measured macroscale material properties. Kohonen maps [61, 62], or self-organizing maps (SOMs), were applied to the dataset in order to conduct a sensitivity analysis of all of these factors and responses. In addition, principal component analysis (PCA) [63] was used to provide a two-dimensional (2-D) representation of nanocomposite data. This facilitated application of the fuzzy C-means (FCM) clustering algorithm [64, 65] to characterize the physical/mechanical properties of VGCNF/VE nanocomposites.

3.1 Materials and Methods

A brief summary of the statistical experimental design and testing procedures to generate the VGFCNF/VE dataset is given here. A more detailed discussion can be found in [10-12, 66].

3.1.1 Statistical Experimental Design

The effect of five input design factors on the viscoelastic properties (storage and loss modulus) of VGCNF/VE nanocomposites were investigated using a general mixed-level full factorial experimental design [67]. These carefully selected factors, based on the state-of-the-art formulation and processing procedures, included: 1) VGCNF type (designated as A), 2) use of a dispersing agent (B), 3) mixing method (C), 4) VGCNF weight fraction in parts per hundred parts of resin (phr) (D), and 5) the temperature (E) used in dynamic mechanical analysis (DMA) testing. Experimental design factors and their associated levels are given in Table 3.1.

Table 3.1 The experimental design factors and their levels [10, 66].

Factor designation	Factors	Level				
		1	2	3	4	5
A	VGCNF type	Pristine	Oxidized	-	-	-
B	Use of dispersing agent	Yes	No	-	-	-
C	Mixing method	US[a]	HS[b]	HS/US	-	-
D	VGCNF weight fraction (phr[c])	0.00	0.25	0.50	0.75	1.00
E	Temperature (°C)	30°C	60°C	90°C	120°C	-

[a] Ultrasonication
[b] High-shear mixing
[c] Parts per hundred parts of resin

A total of 2×2×3×5×4 = 240 "treatment combinations" (different combinations of the factor levels in Table 3.1) were randomized to eliminate bias in preparing the

specimens. Each treatment combination resulted in three specimens prepared from the same material batch [11, 66]. Each specimen was tested using a dynamic mechanical analyzer (single cantilever/flexure mode) to measure average storage modulus, loss modulus, and tan delta for each treatment combination. Storage and loss moduli are dynamic mechanical properties and indicative of the polymer nanocomposite's stiffness and energy dissipation capability, respectively.

3.1.2 Materials and Processing

A low styrene content (33 wt%) VE resin (Ashland Co., Derakane 441-400) and two VGCNF commercial grades, i.e., pristine PR-24-XT-LHT and surface-oxidized 24-XT-LHT-OX, both of them were bought from Applied Sciences, Incorporation. were utilized for nanocomposite specimen preparation [11, 66]. In addition, methyl ethyl ketone peroxide (MEKP) (US Composites Inc.) and 6% cobalt naphthenate (CoNaph) (North American Composites Co.) were selected as initiator and crosslinking promoter, respectively. Air release additives BYK-A 515 and BYK-A 555 (BYK Chemie GmbH) were used to remove air bubbles introduced during mixing. A commercial dispersing agent BYK-9076 (BYK-Chemie GmbH) was employed to improve VGCNF dispersion in the resin.

From a group of resin comprising 100 parts resin, 0.20 phr 6% CoNaph, 0.20 phr BYK-A 515, 0.20 phr BYK-A 555, 0.00-1.00 phr VGCNFs (based on the design given in Table 3.1), and a 1:1 ratio of BYK-9076 to VGCNFs, test specimens were fabricated. The VGCNF/resin blend was mixed by either an ultrasonicator whose model is GEX750-5C from Geneq Incorporation, high shear mixer whose model is L4RT-A from Silverson Machines Ltd., or a combination of both, as dictated by the design given in Table 3.1.

Then the nanofiber/resin blend was degassed under vacuum for 5-15 min at pressures of 8-10 kPa. The blend was thermally cured for 5 h at 60˚C followed by 2 h post-curing at 120˚C.

3.1.3 Dynamic Mechanical Analysis (DMA)

Test specimens were cut from cured specimens for DMA and polished using sandpaper. The storage and loss moduli were measured over a temperature range of 27-160^{o}C using a dynamic mechanical analyzer (TA Instruments, Model Q800) in the single cantilever mode at an amplitude of 15 μm, a fixed frequency of 10 Hz, and a heating rate of 5^{o}C/min.

3.2 Theory/ Calculation

The average storage and loss moduli from three repeat tests for each of the 240 treatment combinations are given in [10]. This study incorporates five input design factors, i.e., VGCNF type (A), use of a dispersing agent or not (B), mixing method (C), VGCNF weight fraction (D), and DMA testing temperature (E) and three output responses, i.e., storage modulus, loss modulus, and tan delta. Hence, the dataset represents an eight-dimensional (8-D) space for analysis. Since factors A, B, and C are considered qualitative factors, they are represented by a numeric code for analysis purposes. For two-level factors A and B, 0 and 1 are the coded values for the first and second levels, respectively. For the three-level factor C, -1, 0, and 1 are the coded values for the first, second, and third levels, respectively (Table 3.1).

The logic behind data mining can be summarized as follows: 1) identify dominant patterns and trends in the data by utilizing the SOMs to conduct a sensitivity analysis; 2)

apply a dimensionality reduction technique, such as PCA, to the data in order to enable the FCM clustering analysis of the data; 3) perform the FCM analysis of the data; and 4) transfer the findings of data mining techniques to the domain experts to validate the discovered data patterns and trends.

On the basis of the above discussion, SOMs [61, 62], PCA [63], and the FCM clustering algorithm [64, 65] were used with the 240 treatment combination dataset to discover nanocomposite data patterns and trends and to identify the different system features related to the specific material properties. SOMs were created with respect to temperature, VGCNF weight fraction, storage modulus, loss modulus, and tan delta. After analyzing the SOMs, temperature was identified as the most important input feature for the VGCNF/VE nanocomposites because it has the highest impact on the resulting storage and loss moduli responses. VGCNF weight fraction was also an important feature. In addition, it was inferred from the SOMs that some specimens tested at the same temperature tended to have several sub-clusters (groups). Each sub-cluster had the same tan delta or VGCNF weight fraction values.

Before applying these techniques, a brief explanation of ANN and unsupervised learning is presented.

3.2.1 Artificial Neural Networks (ANNs) and Unsupervised Learning

ANNs are an architecture of processors or neurons that are connected, organized, and associated with a learning algorithm that emulates a biological process [61]. The interconnections have “weight” values that are adjusted over time to emulate learning. These weights encode knowledge about the problem domain. The neurons and their

connections form a structure that resembles a biological neural network, so different schemes are formulated based on the neuroscience areas [62].

There are two different learning methodologies in any ANN model; supervised or unsupervised [61]. A supervised approach implements a learning algorithm that creates an input to output mapping based on a training dataset whose samples (vectors) are labeled; thus, supervised learning creates a mapping between a space of particular dimensions (n) (input) to a space of particular dimensions (m) (output). Thus, the ANN will learn a relation (function) from the input-output mappings and will have the ability to generalize well such that when a new input vector is provided as input, the network will recognize it and classify it into the corresponding output vector. In contrast, an unsupervised approach uses only a set of unlabeled vectors the ANN must use as inputs and from which it must learn. In other words, the unsupervised ANN is expected to create characterizations about the input vectors and to produce outputs corresponding to a learned characterization (i.e., knowledge discovery).

ANNs that use unsupervised learning will extract clusters or groups based on the similarity of some features of the input dataset so that the results will be presented in a clear and understandable manner [61]. Domain experts must be consulted to check the outputs from the ANNs that use the unsupervised learning approach to determine if there are any trends or patterns that can be inferred from the classification results. Adjustments to one or more of the training parameters used to control the learning process should be made if the classification results are not satisfactory. After this step, the network is presented with the inputs again.

3.2.2 Self-Organizing Maps (SOMs)

Kohonen maps are utilized to map patterns of arbitrary dimensionality into 2-D or 3-D arrays of neurons that form the maps [62]. A SOM works as a self-cluster. The basic components of a 2-D SOM for assessing VGCNF/VE feature data are shown in Figure 3.1. The inputs are the dimensions of the dataset being analyzed. Each component of the input vector x is connected to each unit on the SOM through the weight vector w_{ij}. After some training epoches, a mapping between the nanocomposite input data space and the 2-D map of neurons will be created by the SOM. The nanocomposite feature output y_i specifies the relation between the input vector and the weight vector in each processing unit. SOM utilizes a technique for the nonlinear mapping [68] that keeps the high dimensional order of the SOM. That is, if two input vectors are close to one another in the high dimensional space, then these vectors are close to each other on the map.

In Figure 3.1, a trained feature map and its response to a winning output neuron is shown. This neuron is produced when the original training set or similar input vector set is presented to the SOM [61]. This figure is a general illustration to show the logic of the SOM and the ANN techniques. Domain experts will discover new knowledge based on the winning noutput neuron and the characteristics of the whole resulting SOM.

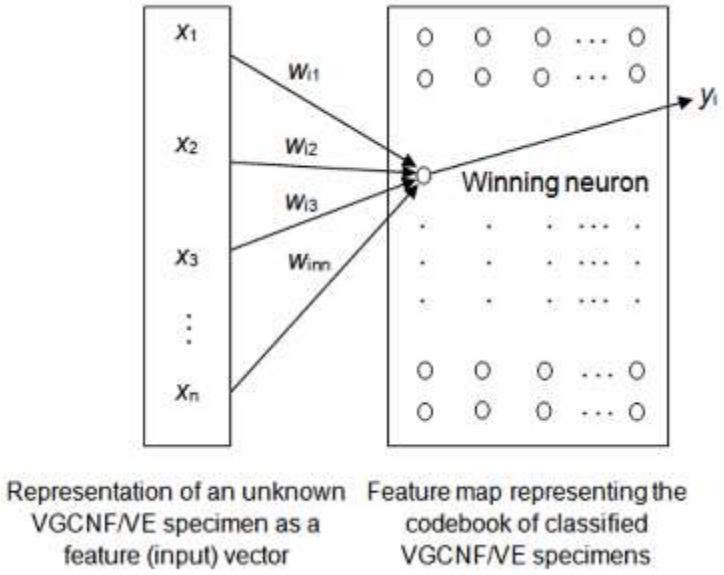


Figure 3.1 Representation of the VGCNF/VE data analysis using ANN and a SOM. In the processing unit, the input vector *x* is multiplied by the weight vector w to create a mapping to the output vector *y*.

The SOM training procedure is implemented on a spatial array of units (neurons) as shown in Figure 3.2 with spatially neighborhoods (hexagonal or rectangular arrays). In addition, the SOM must contain a method of representing a compressed and manageable form of the data. That is, SOM performs the compression without losing information regarding the distance between the corresponding data vectors. Euclidean distance is used by SOM to determine the similarity of data vectors [14].

The spatial neighborhood, N_m, is used in determining the similarity between the input data vector and the vector that represent the weights between the input layer and all the neurons on the SOM. Initially, random values of the weights, learning rate, and neighborhood size are selected. Then, when a training data vector is presented to the network as an input vector, the neuron with the most similar weight values is extracted. Then, the weights of the winning neuron and the neighborhood neurons are adjusted so

that they will be close to the training data vector. While the training process continues, the learning rate values and the neighborhood size should be decreased until no significant adjustments are made by the SOM. The result is that the neurons in the winning neighborhood will be finalized at the current learning step while the weights values in the other neurons remain unchanged. The winning neighborhood is defined as the neighborhood located around *m* which represents the best matching neuron [62].

The SOM algorithm proceeds as follows. First, for every neuron *i* on the SOM, there is a corresponding parametric weight vector w_i. $w_i(0)$ initial values are chosen randomly. Next, an input vector *x* (R_n) is applied simultaneously to all of the neurons on the map. The best-matching neuron is the one with the smallest Euclidean distance. However, other distance measures can be used and compared to the Euclidean distance to determine which measure is the most efficient in clustering the data vectors [61]. As the training process proceeds forward, the radius of N_m decreases with time (*t*) according to $N_{m(t_1)} > N_{m(t_2)} > N_{m(t_3)} > ... > N_{m(t_n)}$, where $t_1 < t_2 < t_3 < ... < t_n$. In other words, the winning neighborhood can contain many neurons when training starts, but at the end of the learning process, the neighborhood contains only the winning neuron. The learning rate of the SOM algorithm decreases with time *t*.

The self-organization aspect of the SOM proceeds as follows: 1) the SOM is provided with a sufficient number of training vectors as inputs; 2) in the winning neighborhood, weights are adjusted in order to reach to an acceptable match; 3) based on the activation received by each neuron, the corresponding weights are adjusted. As a result, this adjustment will make the SOM to respond better to a similar subsequent input pattern. Consequently, the SOM is obtained with weights representing the stationary

probability density function of the data vectors used in the training process. In addition, the SOM displays the data from a different view. That is, instead of viewing the data as an *n*-dimensional vector, it can be displayed as a 2-D representation. By looking at the location of the input vector that represent a sample on the SOM, the data insights and trends as well as the human analysis of the corresponding data can be enhanced because the *n*-dimensional input vector can be represented by a simple but yet reliable SOM 2-D structure [62].

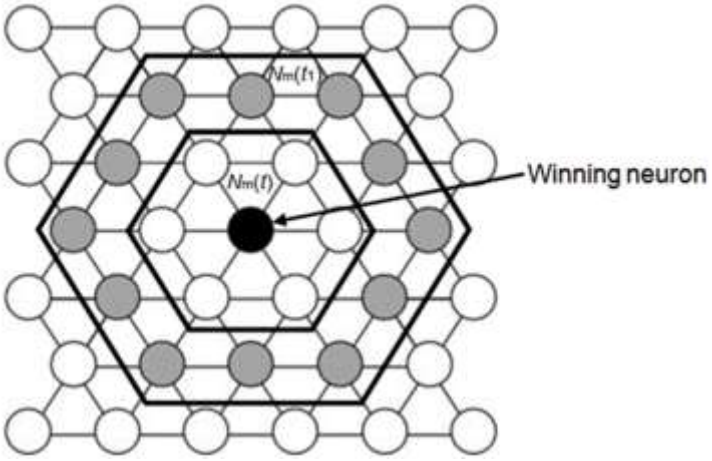


Figure 3.2 Hexagonal four nearest neighbors SOM grid

3.2.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a method of identifying patterns in data and expressing this data to highlight similarities and differences [63]. These patterns can be hard to find in data of higher dimensions, where visual representations are not available. Therefore, PCA can be used as a powerful tool for analyzing data, identifying patterns, and data compression.

After performing PCA, the number of dimensions will be reduced without much loss of the embedded information. PCA includes four main data processing steps. First, the mean, i.e., the average across each dimension, is calculated. Second, the mean is subtracted from each of the data dimensions. Third, the covariance matrix [63] is calculated along with its eigenvalues and eigenvectors. Finally, these eigenvectors and eigenvalues can be used to choose the principal components and form a feature vector in order to derive the new low-dimensional dataset.

3.2.4 Fuzzy C-means Clustering Algorithm

Once data dimensions have been reduced to a 2-D or 3-D graphical representation via PCA, several clustering algorithms can be applied to discover patterns in the data. In the following section, a summary of the FCM clustering algorithm, developed by Bezdek and Ehrlich [65] is presented. Clustering is often associated with the "membership" matrix U [65], which specifies the degree by which a certain data vector x belongs to a particular cluster c. The size of U is $C \times N$, where C is the number of clusters and N is the number of data vectors in the dataset. C is set initially to be $2 \le C \le (N-1)$.

$$U = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1N} \\ u_{21} & u_{22} & \cdots & u_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ u_{C1} & u_{C2} & \cdots & u_{CN} \end{bmatrix} \tag{3.1}$$

$$\text{where } u_{ij} = \begin{cases} 1 & if\ x_j \in A_i \\ 0 & otherwise \end{cases} \tag{3.2}$$

u_{ij} is called a crisp 0-1 matrix and x_j and A_i represent the data vector j and the class i, respectively. The number of elements in a cluster is given by the sum across a row of U, and

$$\sum_{i=1}^{C} u_{ij} = 1 \quad \text{for all } j = 1, 2, \ldots, N, \tag{3.3}$$

The membership matrix U can be classified into three different categories based on its values:

- Possibilistic:

$$N_{pc} = \left\{ p \in \Re^c : p_i \in [0,1], p_i > 0 \; \exists_i \right\} \tag{3.4}$$

- Fuzzy or probabilistic:

$$N_{fc} = \left\{ p \in N_{pc} : \sum_{i=1}^{C} p_i = 1 \right\} \tag{3.5}$$

- Crisp:

$$N_{hc} = \left\{ p \in N_{fc} : p_i \in \{0,1\} \right\} \tag{3.6}$$

Here, N_{hc} is the canonical (i.e., unit vector) basis of $\Re^c$. The ith vertex of N_{hc}, i.e.,

$$e_i = \left(0,0,...., \underbrace{1}_{ith\ place}, 0,....,0 \right)^T \tag{3.7}$$

Where the 1 occupies the *i*th place, is the crisp label for class *i*, $1 \le i \le c$. The set N_{fc} is a piece of a hyperplane, and is the convex hull of N_{hc}. For example, the vector $p = (0.1, 0.6, 0.3)^T$ is a label vector in N_{f3}; its entries lie between 0 and 1, and sum equals 1.

Clustering can be described using an optimization scheme, which involves formulating a cost function and then using iterative and alternate estimations of the function. For example, the cluster centers and membership matrix U can be initially computed and then iteratively recalculated and updated.

FCM was created by Bezdek and Ehrlich [65] and is considered an objective function- based clustering technique. Each cluster using FCM has a prototype v_i that distinguishes cluster *i*, where the initial values of v_i can be set randomly or by picking the furthest points in the dataset or by picking exemplars from the dataset. Thus, the overall prototype vector V has a size of $(1 \times C)$ and can be denoted as

$$V = \{v_1, v_2, ..., v_c\} \tag{3.8}$$

The FCM cost function can be written as

$$J(U,V) = \sum_{i=1}^{C} \sum_{k=1}^{N} u_{ik}^{Q} d(x_k, v_i) \tag{3.9}$$

where Q is a weighting exponent ($1 \le Q < \infty$) and $d(x_k, v_i)$ is the distance measure between the data vector x_k and the cluster *i* (represented by prototype *i*). So,

$$J(U,V)=\sum_{i=1}^{C}\sum_{k=1}^{N}u_{ik}^{Q}d(x_k,v_i)-\sum_{k=1}^{N}\lambda_k\left(\sum_{i=1}^{C}u_{ik}-1\right) \tag{3.10}$$

$$\frac{\partial J(U,V)}{\partial u_{rs}}=Qu_{rs}^{Q-1}d(x_s,v_r)-\lambda_s=0 \tag{3.11}$$

and

$$u_{rs}=\left(\frac{\lambda_s}{Qd(x_s,v_r)}\right)^{\frac{1}{Q-1}} \tag{3.12}$$

resulting in

$$\sum_{i=1}^{C}\left(\frac{\lambda_s}{Qd(x_s,v_i)}\right)^{\frac{1}{Q-1}}=1 \tag{3.13}$$

and

$$\lambda_s=\frac{Q}{\left(\sum_{i=1}^{C}\left(\frac{1}{d(x_s,v_i)}\right)^{\frac{1}{Q-1}}\right)^{Q-1}} \tag{3.14}$$

After substituting λ_s back into $\frac{\partial J(U,V)}{\partial u_{rs}}=0$,

$$u_{rs}=\frac{1}{\sum_{i=1}^{C}\left(\frac{d(x_s,v_r)}{d(x_s,v_i)}\right)^{\frac{1}{Q-1}}} \tag{3.15}$$

After deriving the cost function *J* with respect to prototype *v*,

$$\frac{\partial J(U,V)}{\partial v_j}=\sum_{k=1}^{N}u_{jk}^{Q}\frac{\partial d(x_k,v_j)}{\partial v_j}=0 \tag{3.16}$$

For the Euclidean distance measure,

$$d_{ik}^2(x_k, v_i) = d_{ik}^2 = (x_k - v_i)^T (x_k - v_i) = x_k{}^T x_k - 2x_k{}^T v_i + v_i{}^T v_i \tag{3.17}$$

$$\frac{\partial d^2(x_k, v_i)}{\partial v_i} = -2x_k + 2v_i = -2(x_k - v_i) \tag{3.18}$$

Then

$$\frac{\partial J(U,V)}{\partial v_i} = -2\sum_{k=1}^{N} \left(u_{jk}\right)^Q (x_k - v_j) = \sum_{k=1}^{N} \left(u_{jk}\right)^Q x_k - v_j \sum_{k=1}^{N} \left(u_{jk}\right)^Q = \vec{0} \tag{3.19}$$

Therefore,

$$v_j = \frac{\sum_{k=1}^{N} \left(u_{jk}\right)^Q x_k}{\sum_{k=1}^{N} \left(u_{jk}\right)^Q} \tag{3.20}$$

Now, for the Gustafon-Kessel (GK) distance measure,

$$d_{ik} = \left(\left|\Sigma_i\right|^{\frac{1}{D}} \left(\left(x_k - v_i\right)^T \Sigma_i{}^{-1} \left(x_k - v_i\right) \right) \right)^{\frac{1}{2}} \tag{3.21}$$

As can be seen, d_{ik} is scaled by the hyper-volume approximation denoted by $\left|\Sigma_i\right|^{\frac{1}{D}}$

where Σ_i is the covariance matrix for class *i*:

$$\frac{\partial d_{ik}{}^2}{\partial v_i} = -2\left|\Sigma_i\right|^{\frac{1}{D}} \Sigma_i{}^{-1} (x_k - v_i) \tag{3.22}$$

Therefore,

$$v_i = \frac{\sum_{k=1}^{N} (u_{ik})^Q x_k}{\sum_{k=1}^{N} (u_{ik})^Q} \tag{3.23}$$

$$\Sigma_i = \frac{\sum_{k=1}^{N} (u_{ik})^Q (x_k - v_i)(x_k - v_i)^T}{\sum_{k=1}^{N} (u_{ik})^Q} \tag{3.24}$$

The GK distance measure in Equation 3.21 uses a cluster-specific covariance matrix, so as to adapt various sizes and forms of the clusters. Thus, clustering algorithms that utilize GK distance measures try to extract much more information from the data than the algorithms based on the Euclidean distance measure [65]. Hence, the GK distance measure was used in this study. Based on this development, the *pseudo code* of the FCM algorithm is given as follows:

Compute $C \times N$ distance matrix;

Choose $v_j(0)$ as initial estimates of v_j, $j = 1, \ldots, C$;

//Initial value of the iteration counter, t

$t = 0$;

//Update the membership matrix U

Repeat:

for $i = 1$ to N

for $j = 1$ to C

$$u_{ji} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d(x_i, v_j)}{d(x_i, v_k)} \right)^{\frac{1}{Q-1}}} \quad ;$$

End for

End for

//Now, $t = 1$

$t = t + 1$;

//Prototypes Update

for $j = 1$ to C

solve:

$\sum_{i=1}^{N} u_{ji}^{Q}(t-1) \frac{\partial d(x_i, v_j)}{\partial v_j} = 0$; with respect to v_j and set v_j equal to the computed solution

End for

- Test for convergence:

  Select termination criteria using, for example, particular number of iterations or the difference from t to t-1 of the sum of prototype differences or other appropriate criteria.

3.2.5 K-means Clustering Algorithm

K-means clustering [69] is a form of unsupervised learning whereby a set of observations (i.e., data points) is partitioned into natural grouping or clusters in such a way that the measure of similarity for the observations assigned to each cluster minimizes a specified cost function.

Let $\{x_i\}_{i=1}^{N}$ denote a set of multidimensional observations that is to be partitioned into *K* clusters and let $j = C(i), \quad i = 1,2,...,N, \quad j = 1,2,...,K, \quad and \ K << N$

Denote the partition that assigns the *i*th observation $\boldsymbol{x_i}$ into the *j*th cluster. The cost function for such a partition is:

$$J(C) = \sum_{j=1}^{K} \sum_{C(i)=j} \left\| x_i - \hat{\mu}_j \right\|^2 \tag{3.25}$$

where $\hat{\mu}_j$ is the cluster center of the *j*th cluster. A partition *C* that minimizes this cost is the one that cluster centers can well represent the assigned data points in terms of average similarity.

The following are the steps used in iterative *k*-means clustering algorithm [69]:

- Initiating the centers: $\hat{\mu}_j^{(0)}, j = 1,2,...,K$
- Similarity matching:

  compute the distance (similarity) of a data point to each of the centers, and assign it to the cluster to which it has the minimum distance.

- Updating:

  use the mean of data points temporarily assigned to each cluster to update cluster centers, resulting in $\hat{\mu}_j^{(1)}, j = 1,2,...,K$

- The similarity matching and center updating are continued until there is no data point being shuffled from one cluster to another or $\hat{\mu}_j^{(n)}, j = 1,2,...,K$ do not change after the *n*th iteration.

Note: there are different metrics that can be used to measure the distance such as, Euclidean distance (L2), L1 distance, Mahalanobis distance, spectral angle, and correlation coefficient.

3.2.6 Support Vector Machines (SVMs)

A support vector machine (SVMs) [70] is one of the most robust and widely used classifiers. This technique uses a theory that has been validated from datasets of different sizes and dimensions and from different fields and domains.

SVMs can be used for supervised and unsupervised learning problems. The former, ideally, requires large numbers of data vectors (points) within particular dataset in order for the SVM to generalize well and avoid the over-training (over-fitting) trap for which many knowledge discovery algorithms are vulnerable. The latter, however, can be used and utilized with less data vectors, but there must be some insights in the dataset that assist the SVM model to generalize and predict the correct quantity given unseen data vector [70].

SVMs can be used for designing a classifier that classifies linearly and nonlinearly separable data into two or more classes.

The basic idea of the SVM classifier is to find a separating hyperplane between the points that belong to two classes such that the distance between the closest points (of each class) to this hyperplane is maximized. That is, SVMs techniques seek the availability of the maximum margin that defines the optimal separating hyperplane.

The basic idea of SVM is illustrated in Figure 3.3.

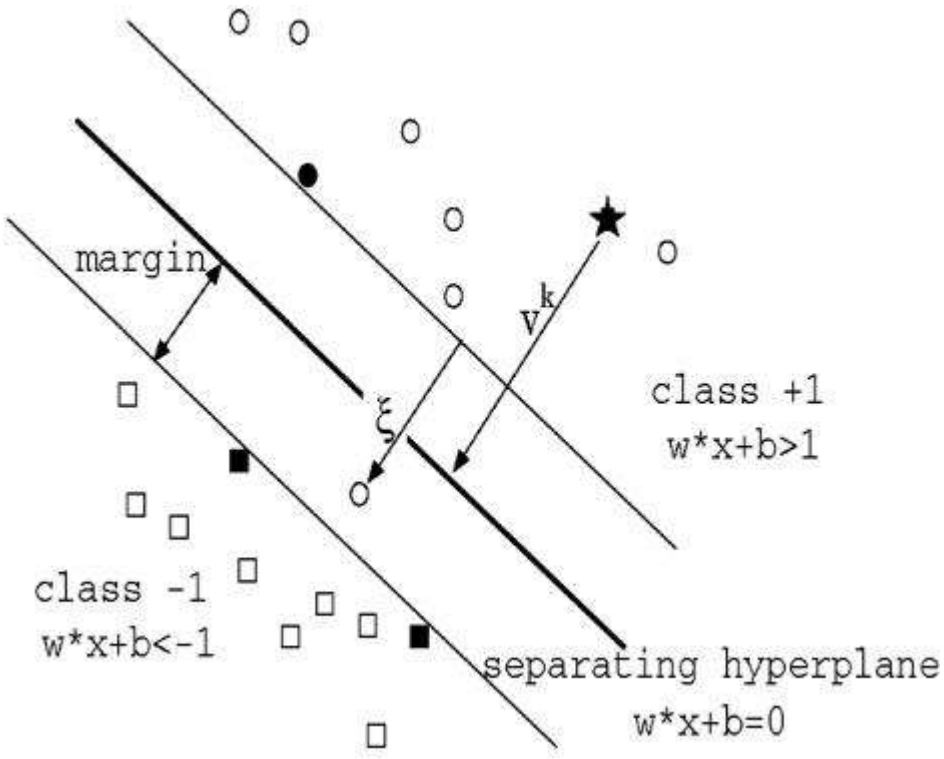


Figure 3.3 The SVM model

The separating hyperplane (that separates two classes) along with the maximum margin that defines the optimal separating hyperplane can be clearly seen.

The margin (m) (indicated in Figure 3.3) is given by the relation:

$$m = \frac{|g(x)|}{\|w\|} \tag{3.26}$$

Where *g(x)* is the discriminant function used to separate the classify data vectors to the corresponding classes. *w* is the weight vector used by SVM model. The main idea here is to scale *w* so that the value of *g(x)* at the closes point to the separating hyperplane is equal to 1 for class 1 and -1 for class 2.

Alternatively, in the case of two classes SVM model, the goal is having a margin such that:

$$m = \frac{1}{\|w\|} + \frac{1}{\|w\|} = \frac{2}{\|w\|} \tag{3.27}$$

and this requires that:

$$\begin{aligned} w^T x + w_0 &\geq 1 \ldots \ \ for \ \ x \in class1 \\ w^T x + w_0 &\leq -1 \ldots for \ \ x \in class2 \end{aligned} \tag{3.28}$$

Where w_0 is the weight bias used to specify how much the margin is away from the origin. *x* is the data point (data vector).

The support vectors λ values are the data points which are located in the margin borders and their values will be greater than zero. λ values that are less than zero are not considered support vectors and hence the corresponding data points belong to either class 1 or class 2.

The theory of SVMs states that for each data vector x_i, there must be a class indicator (say y_i). As mentioned above, 1 can be specified for data vectors that belong to class 1 and -1 for data vectors that belong to class 2. The task is to find *w*, and w_0 such that the cost function (Equation 3.25) is minimized.

$$J(w, w_0) = \frac{1}{2}\|w\|^2 \tag{3.29}$$

Subject to:

$$y_i(w^T x + w_0) \geq 1 \quad for \;\; i = 1,2,...,N \tag{3.30}$$

This nonlinear optimization task can be solved using a quadratic programming which is considered as an optimization algorithm to maximize a quadratic function of some real-valued variables subject to linear constraints [71].

In SVMs, a Lagrangian function can be used. That is,

$$L_p(w, w_0, \lambda) = \frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i \left[y_i (w^T x_i + w_0) - 1 \right] \tag{3.31}$$

where λ is the Lagrange multiplier vector and λ_i is the Lagrangian multiplier. $L_p(w, w_0, \lambda)$ in this case is usually referred to as *Lagrangian in primal form*.

The Lagrangian function is subject to a set of constraints and the Karush-Kuhn-Tucker (KKT) conditions define these constraints. The KKT conditions are:

- $\frac{\partial}{\partial w} L(w, w_0, \lambda) = 0$
- $\frac{\partial}{\partial w_0} L(w, w_0, \lambda) = 0$
- $\lambda_i \geq 0$ for i=1,2,….,N
- $\lambda_i \left[y_i (w^T x_i + w_0) - 1 \right]$ for i =1,2,….,N

Now, if Lagrangian function equation $L(w, w_0, \lambda)$ is combined with the first and second KKT conditions, the SVM optimization task is to minimize $L(w, w_0, \lambda)$, subject to

the following constraints; (*Note:* the first two are the equality constraints and the last one is the inequality constraint)

$$w = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{3.32}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{3.33}$$

$$\lambda_i \geq 0 \tag{3.34}$$

If the equalities above are substituted into $L(w, w_0, \lambda)$, then the final form of SVM optimization task for the two-class linearly separable case is to minimize Equation (3.35) with respect to λ.

$$L_D(w, w_0, \lambda) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j \tag{3.35}$$

Subject to the following constraints:

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

$$\lambda_i \geq 0$$

$L_D(w, w_0, \lambda)$ is usually referred to as *Lagrangian in dual form*.

The weight bias term can be calculated from the relations above. For example, if $\lambda_1 > 0$, then the corresponding w_0 can be found from the relation:

$$y_1(w^T x_1 + w_0) = 1 \tag{3.36}$$

Therefore,

$$w_0 = \frac{1}{y_1} - w^T x_1 \tag{3.37}$$

Figure 3.4 illustrates a visualization of the SVM implementation of nonlinearly separable data.

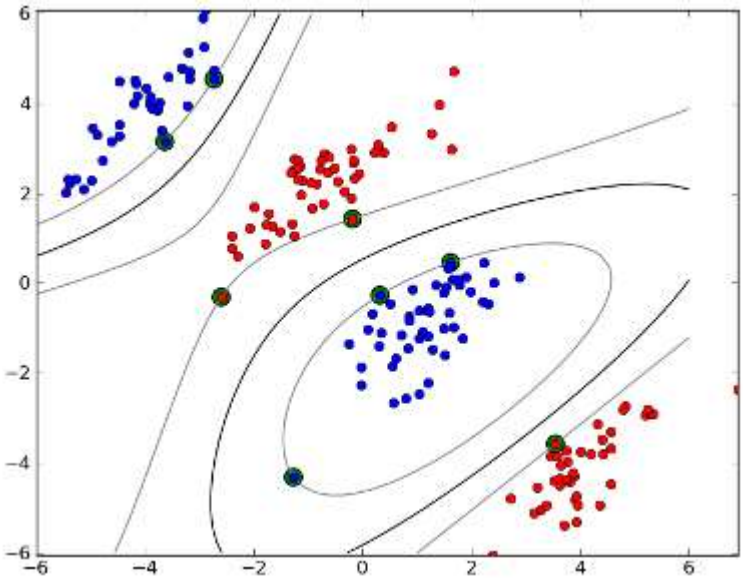


Figure 3.4 An example of nonlinearly separable data.

This case involves introducing a new set of "slack" variables (ξ)

The basic idea of SVM implantation for the non-linearly separable data is that a new set of "slack" variables are introduced such that:

$$y_i[w^T x + w_0] \geq 1 - \xi_i \tag{3.38}$$

In this context, the following scenarios must be taken into account:

- correct classification of the data point x_i is obtained if $\xi_i = 0$

- x_i will be inside the band (inside the margin) if $0 \le \xi_i \le 1$
- x_i is misclassified (the SVM model classify it in a different class than what it actually should belong) if $\xi_i > 1$

The closely related cost function in this case (in *primal form*) is to minimize $J(w, w_0, \xi)$ such that:

$$J(w, w_0, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \tag{3.39}$$

Subject to the constraints:

$$y_i[w^T x + w_0] \ge 1 - \xi_i \tag{3.40}$$

$$\xi_i \ge 0 \tag{3.41}$$

for $i = 1,2,\ldots,N$

where C is a positive constant that balances between the margin size and the misclassification instances. The choice for C determines the number of support vectors and the overall performance of the SVM model.

The corresponding Lagrangian function in this case becomes as:

$$L(w, w_0, \xi, \lambda, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} \lambda_i [y_i (w^T x_i + w_0) - 1 + \xi_i] \tag{3.42}$$

where λ and μ are the Lagrangian vectors. The corresponding KKT conditions in this case are:

- $\frac{\partial}{\partial w} L(w, w_0, \xi, \lambda, \mu) = 0$

- $\frac{\partial}{\partial \xi_i} L(w, w_0, \xi, \lambda, \mu) = 0$ $\frac{\partial}{\partial w_0} L(w, w_0, \xi, \lambda, \mu) = 0$
- $\lambda_i \geq 0$ for i=1,2,….,N
- $\mu_i \geq 0$ for i=1,2,….,N
- $\mu_i \xi_i = 0$ for i=1,2,….,N
- $\lambda_i [y_i (w^T x_i + w_0) - 1 + \xi_i]$ for i=1,2,….,N

The goal in non-linearly separable case is to make the margin as large as possible but at the same time the number of data points with $\xi > 0$ as small as possible. In this case, the misclassifications mistakes as well as the case where there are some data points inside the margin even though it's correctly classified will be avoided.

Therefore, we the Lagrangian function $L(w, w_0, \xi, \lambda, \mu)$ (in *dual form*) can be minimized subject to the following constraints:

$$w = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{3.43}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{3.44}$$

$$C - \mu_i - \lambda_i = 0$$
$$\lambda_i \geq 0,\ \mu_i \geq 0 \tag{3.45}$$

If the above equality constraints are substituted in the Lagrangian, the final *dual form* format of nonlinearly separable data $L(w, w_0, \xi, \lambda, \mu)$ must be minimized with respect to λ such that:

$$L(w, w_0, \xi, \lambda, \mu) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j \tag{3.46}$$

Subject to the constraints:

$$0 \le \lambda_i \le C, \quad for\ i = 1,2,....,N \tag{3.47}$$

$$\sum_{i=1}^{N} \lambda_i y_i = 0 \tag{3.48}$$

Another important aspect of SVMs development is the kernel function which takes the optimization problem from a lower space to a higher space. This kernel function is a function of x_i and x_j shown above for the dual form and so the SVM optimization task becomes to minimize $L(w, w_0, \xi, \lambda, \mu)$ with respect to λ such that:

$$L(w, w_0, \xi, \lambda, \mu) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j K(x_i, x_j) \tag{3.49}$$

Subject to the constraints in Equations 3.47 and 3.48. $K(x_i,x_j)$ is the implemented kernel function.

The following are some of the typical kernels:

- Polynomials:

$$K(x, z) = (x^T z + 1)^Q \quad Q > 0 \tag{3.50}$$

  where Q is the polynomial degree

- Radial basis functions (RBF)

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \tag{3.51}$$

where σ^2 is the standard deviation

- Hyperbolic Tangent:

$$K(x,z) = \tanh(\beta x^T z + \gamma) \tag{3.52}$$

where β, γ are constants

- Dot product:

$$K(x,z) = x^T z \tag{3.53}$$

The following strategy to assign each data point (data vector) to the corresponding class can be used:

$$\begin{aligned} g(x) &= \sum_{i=1}^{NS} \lambda_i y_i K(x_i, x) + w_0 > 0, \text{ then } x_i \in class\,1 \\ g(x) &= \sum_{i=1}^{NS} \lambda_i y_i K(x_i, x) + w_0 \le 0, \text{ then } x_i \in class\,2 \end{aligned} \tag{3.54}$$

where $g(x)$ is the discriminant function, *NS*: is the number of support vectors utilized from the problem, and x_i is a data vector in the dataset being optimized.

3.2.7 ANNs Resubstitution Method

The resubstitution method is computationally efficient and requires construction of only one ANN model, which is used for both application and validation [72]. That is, in this method, the whole dataset is used to train the ANN and the same dataset is used for testing (validation). This ensures that the designed ANN model generalizes well and is able to predict the unknown outputs (responses) for new data samples. Good

generalization is achieved when the mean square error (MSE) between the actual responses of ANN model and the desired (targeted) responses is minimal [72].

The MSE is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(e_i)^2 = \frac{1}{N}\sum_{i=1}^{N}(t_i - a_i)^2 \tag{3.55}$$

where N: is the total number of samples,

t_i: is the targeted response, and

a_i: is the actual ANN response

Although several network architectures and training algorithms are available, the feedforward artificial neural network (FFANN) trained by the back-propagation (BP) algorithm is most commonly used, and was used in this dissertation [73]. A FFANN is a layered network of artificially processing units, called neurons, in which connections between neurons are associated with weights that represent the strength of the connections [61]. It includes an input layer, with one neuron corresponding to each of the inputs used in the model, and an output layer, with a single neuron corresponding to each output variable (response). The network also includes one or more ‘‘hidden” layer(s) of neurons. Because each neuron in the hidden layer is associated with an activation function (sigmoidal function in this study), the hidden layer(s) allows a non-linear mapping from the values of the input variables to the value of the output variable [61]. A FFANN network is trained on a dataset of related input-output examples to estimate a non-linear relationship between the input variables and the output variable(s). Upon presenting the training samples to the network, the weighted connections between neurons are adjusted by BP to decrease the MSE between the network’s output and the

targeted output. The process is repeated until the MSE has been reduced as far as possible [74].

3.2.8 Cross Validation Technique

Cross validation (CV) is a technique that can be used to better train the ANN with the available samples in the dataset. First, the available dataset is randomly partitioned into a training set and a test set. The training samples are further partitioned into two disjoint subsets; the estimation subset which is used to select the ANN model and determine the interconnection weights and the validation subset for testing and validating the developed ANN [75].

The essence of CV technique is validating the developed model on a dataset other than the one by which the model's parameters and variables are estimated. Different ANN architectures (frameworks) can be developed and the training dataset can be used to check their performance. Eventually, the architecture that yields the best performance will be adapted [75].

There are four different methods of CV technique and the following is a brief explanation of each of these methods [74].

- Holdout validation: if a random number $r \in [0, 1]$, then (1- r)N samples are allotted to the estimation subset, and the remaining rN samples for validation. This method is computationally expensive and the final ANN model is the one yielding the minimum validation error. When the complexity of the target function (input-output mapping) is small compared with the sample size N, the validation performance is relatively

insensitive to the choice of r, whereas when the target function becomes more complex relative to the sample size N, the choice of r has a more pronounced effect on cross-validation performance. However, a single fixed value of r (e.g., 0.2) works nearly optimally for a wide range of target-function complexity.

- Early-stopping method of training: with good generalization as the goal, the training can be stopped earlier before the learning error becomes too low. The best point to stop training can be determined by the periodic "estimation-followed-by-validation" process as shown in Figure 3.5. After some periods of training, say five epochs, the ANN weights are fixed and the validation error is then measured. The training process is reset for a new epoch(s), when validation process is completed. Finally, when the validation error starts to increase, it is the point to terminate the training process and finalize the weights, even if the error for the training samples continues to decrease.
- Multifold validation: The disadvantage of holdout method is that not all the samples are used for validation. Instead, in multifold validation, the N samples are divided into K subsets. Each time, one subset is used for validation and the remaining K-1 subsets for training. The process is continued until each subset is used for validation once. In this study, 3-folds cross validation was implemented and the performance is assessed by averaging the validation error over all the trials. In Figure 3.6, an illustration of 4-folds cross validation is shown.

- Leave-one-out validation: When the available number of samples, *N*, is severely limited, an extreme form of multifold validation known as leave-one-out validation can be used. In each trial, *N*-1 samples are used for training and the one left out can be used for testing. The process is repeated N times until each sample is used for validation exactly once.

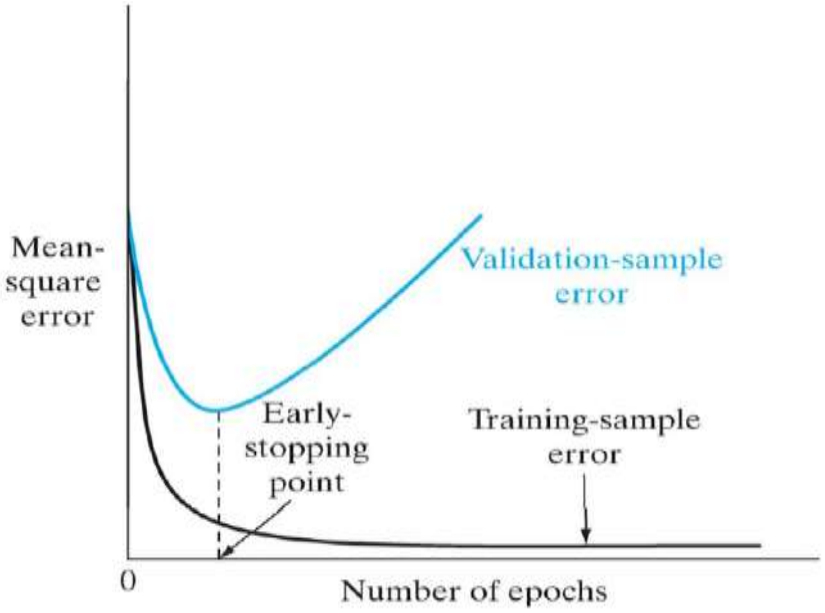


Figure 3.5 A graphical representation of the cross validation technique when early-stopping rule is implemented [75].

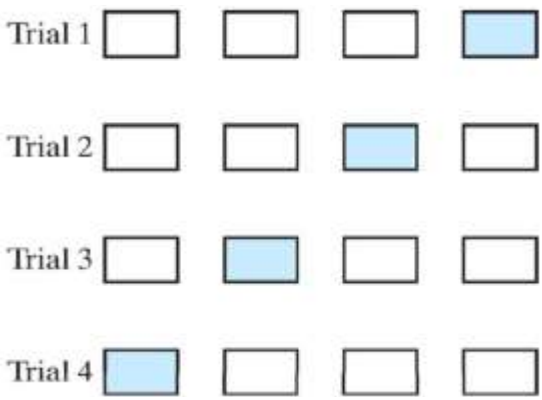


Figure 3.6 Illustration of the multifold method of cross validation.

For a given trial, the subset of shaded data is used to validate the model and the remaining data is used to train the model [75].

CHAPTER IV

RESULTS AND DISCUSSION

In this chapter, the results obtained after applying signal processing, data mining and knowledge discovery algorithms and techniques, outlined in Chapter 3, to VGCNF/VE and AHSS datasets are thoroughly analyzed and discussed. The VGCNF/VE dataset consisted of 240 data points each corresponding to the combinations of five input design factors and three output responses, i.e., a total of eight "dimensions." The AHSS dataset consisted of 19 data points for 1st generation AHSS and 4 data points for 2nd generation AHSS. The dataset dimensions are the ultimate tensile strength (UTS) and the chemical compositions. Once again, for the ANN implementation in this dissertation, following the standard practice of the ANN analysis, the inputs and outputs were normalized using standardized scores, as their original value ranges were completely different from each other whereas for the SOM, PCA, and FCM, the original values were retained as these techniques either analyze the whole dataset, including its inputs and outputs, or analyze one dimension at a time without dividing the dataset into inputs and outputs like in ANN approaches.

4.1 VGCNF/VE Results and Analysis

In Figure 4.1, a 10×10 SOM resulting from the 240 data points is shown. Nanocomposite specimens tested at the same dynamic mechanical analysis (DMA)

temperature tend to cluster together. For example, specimens tested at 30°C tend to cluster at the top of the map, whereas specimens tested at 120°C tend to cluster at the bottom. A mixture of specimens tested at 60°C and 90°C are located in the middle of the map.

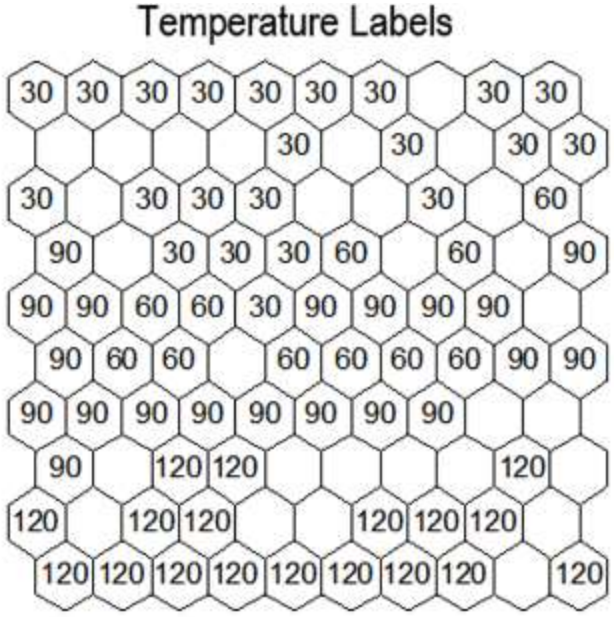


Figure 4.1 A 10×10 SOM with respect to temperature

Note that this is for the 240 nanocomposite specimens used in the study (with all eight dimensions). The specimens tested at the same temperature tend to cluster together.

In order to conduct a complete sensitivity analysis of VGCNF/VE materials system, SOMs are shown and analyzed for other VGCNF/VE features and responses. In Figures 4.2 and 4.3, two 10×10 SOMs for the VGCNF weight fraction and the tan delta (ratio of loss to storage modulus) response are shown, respectively. In Figure 4.2, specimens with the same weight fraction tend to cluster together, but this tendency is not consistent and is less than the clustering tendency shown in Figure 4.1 for temperature.

However, if the cluster at one temperature (say 30°C on the top of Figure 4.1) is considered and compared to the corresponding cluster in Figures 4.2 and 4.3, sub-clusters with similar VGCNF weight fractions can be identified. For example, the first seven 30°C labels from the left in Figure 4.1 have corresponding weight fractions of 0.25, 0.25, 0, 0.50, 1.00, 1.00, and 0.5 phr and corresponding tan delta values of 0.03, 0.02, 0.03, 0.02, 0.02, 0.02, and 0.02. This means that within the nanocomposite specimens tested at 30°C, there are some specimens with similar VGCNF weight fractions that tend to cluster together. For example, specimens with a weight fraction of 0.25 phr as well as 1.00 phr are mapped together (Figure 4.2). Similarly, specimens that have a tan delta value of 0.02 are mapped together (Figure 4.3). This explains why some of the specimens tested at the same temperature are separated by blank hexagons from each other. Each group of specimens in Figures 4.2 and 4.3 that were tested at the same temperature tend to have similar VGCNF weight fractions or tan delta values. However, in Figure 4.3, the clustering for tan delta is more pronounced than that of the VGCNF weight fraction and less than that of the temperature. This leads to the conclusion that temperature is the dominant feature for the treatment combinations and has the highest impact on the responses followed by tan delta and VGCNF weight fraction.

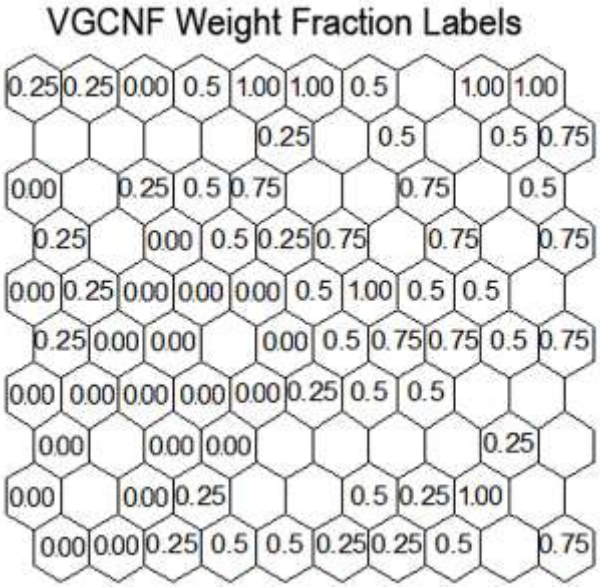


Figure 4.2 A 10×10 SOM with respect to VGCNF weight fractions.

The clustering tendency is less than that of the temperature in Figure 4.1. However, within a certain temperature cluster, the existence of sub-clusters with the same VGCNF weight fraction is confirmed.

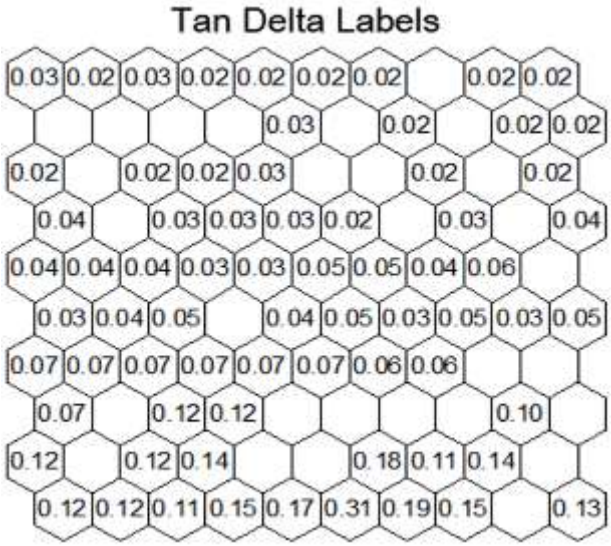


Figure 4.3 A 10×10 SOM with respect to tan delta values.

The clustering tendency is less than that of the temperature in Figure 4.1. However, within a certain temperature cluster, the existence of sub-clusters with the same tan delta value is confirmed.

In addition to the sensitivity analysis inferred from SOMs, the different conditions needed to produce a particular response can also be determined. In Figure 4.4, a 10×10 SOM is shown indicating the indices, which represent the numeric orders of the specimens mapped. Each index corresponds to one treatment combination out of 240 with specific values of VGCNF type, use of a dispersing agent, mixing method, VGCNF weight fraction, testing temperature, storage modulus, loss modulus, and tan delta. The indices in Figure 4.4 can be used to extract information linking the different dimensional combinations that produce certain response values. For example, in Figures 4.5 and 4.6, SOMs for the storage and loss moduli are illustrated, respectively.

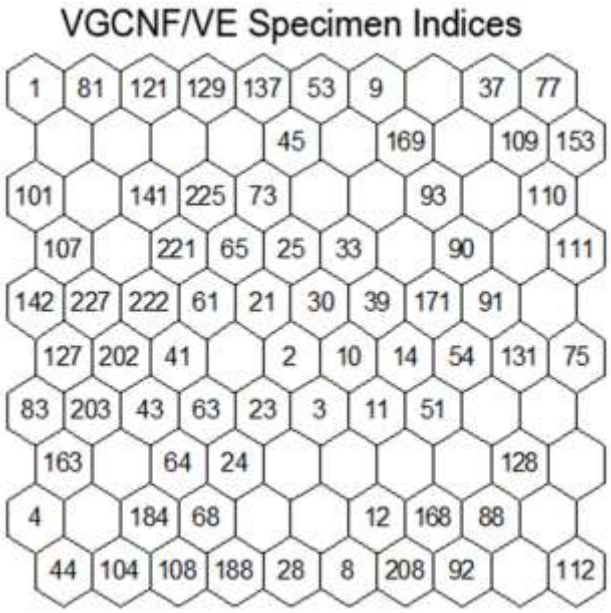


Figure 4.4 A 10×10 SOM illustrating the indices (numeric orders) of the 240 nanocomposite specimens [16].

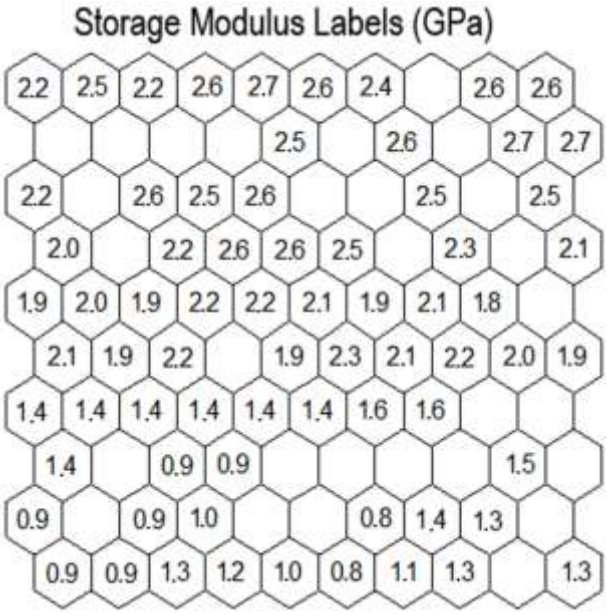


Figure 4.5 A 10×10 SOM based on the storage modulus response.

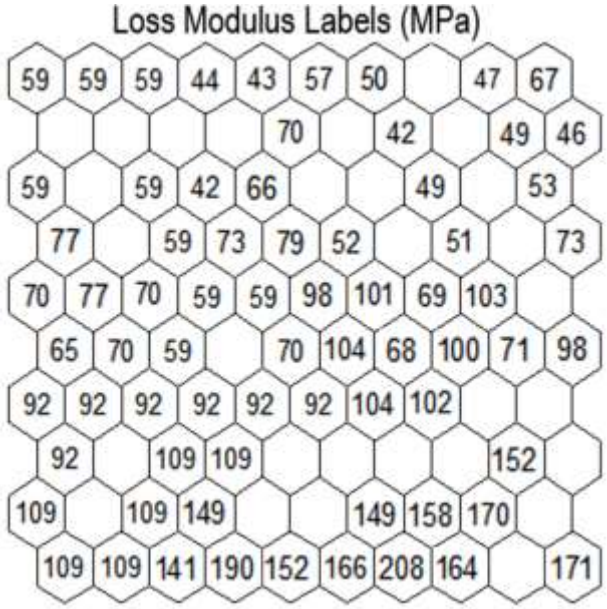


Figure 4.6 A 10×10 SOM based on the loss modulus response

The values are rounded to the nearest integer for simplicity.

In Figure 4.5, the storage modulus response values are shown. A group of three specimens have a storage modulus of about 2.6 GPa, located at the third and fourth rows of the SOM. In Figure 4.4, these values correspond to specimen indices 73, 65, and 25. Clearly, different dimensional properties can be determined to produce the same 2.6 GPa response value. These properties are shown in Table 4.1, where the third row (highlighted) has a lower tan delta value and higher storage modulus than the other two specimens. Nanocomposite designers can use such information in the selection of input factor levels.

Table 4.1 Different dimensional combinations required to produce a storage modulus of about 2.6 GPa.

VGCNF Type (A)	Use of a Dispersing Agent (B)	Mixing Method (C)	VGCNF Weight Fraction (D) (phr)	Temperature (E) (°C)	Storage Modulus (GPa)	Loss Modulus (MPa)	Tan Delta
Pristine	Yes	US[1]	0.25	30	2.577	79	0.031
Oxidized	Yes	US	0.25	30	2.566	73	0.028
Oxidized	Yes	US	0.75	30	2.641	66	0.025

[1] Ultrasonication

Similarly, the loss modulus responses for a group of three specimens are all about 104 MPa in the sixth and seventh rows of the SOM in Figure 4.6. These correspond to indices 10, 11, and 51 (Figure 4.4). Again, different dimensional properties can be prescribed to produce the 104 MPa responses. These properties are shown in Table 4.2, where the first row (highlighted) has a lower tan delta value and higher storage modulus response than the other two specimens.

Table 4.2 Different dimensional combinations required to produce a loss modulus of about 104 MPa.

VGCNF Type (A)	Use of a Dispersing Agent (B)	Mixing Method (C)	VGCNF Weight Fraction (D) (phr)	Temperature (E) (°C)	Storage Modulus (GPa)	Loss Modulus (MPa)	Tan Delta
Pristine	No	US[1]	0.5	60	2.276	104	0.046
Pristine	No	US	0.5	90	1.621	104	0.064
Oxidized	No	US	0.5	90	1.614	102	0.063

[1] Ultrasonication

A PCA was run on the VGCNF/VE nanocomposite data. Figure 4.7 shows a graphical representation for the PCA of the data. PCA reduced the number of data dimensions from eight to two and each specimen was a given a specific 2-D representation (principal component 1 and 2 axes) so that specimens that have similar properties were mapped together in the 2-D space. Thus, there are no specific units associated with the abscissa and ordinate. This step is fundamental so that clustering algorithms (Sections 4.2.4 and 4.2.5) can be applied to identify certain patterns in these nanocomposite data. Such patterns can be used to explain certain physical/mechanical behavior associated with the data without running additional experiments.

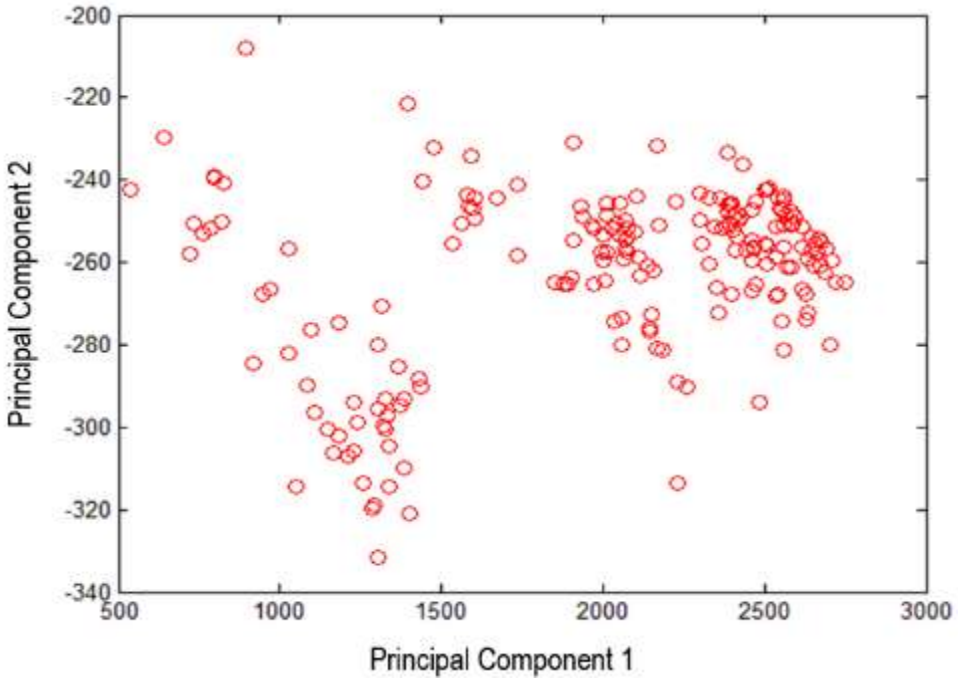


Figure 4.7 A 2-D graphical representation of the VGCNF/VE nanocomposite specimen data (illustrated by circle points) using the PCA technique.

This technique maps the data from an 8-D space down to a 2-D space so that different clustering algorithms can be applied. The values associated with the principal dimensions 1 and 2 are random, but each specimen was given a 2-D coordinate so that specimens with similar properties would be mapped together in the 2-D space.

The FCM was applied to the VGCNF/VE nanocomposite data using the GK distance measures. In Figure 4.8, the FCM results are illustrated, where four clusters are chosen to represent the data using the GK distance measure. In Figure 4.8a, the data points are divided into four different clusters, each shown with a different color. In Figure 4.8a, the nanocomposite specimens tested at 90^{o}C and 120^{o}C each form a separate cluster with average tan delta values of 0.049 and 0.148, respectively. The rest of the nanocomposite specimens tested at 30^{o}C and 60^{o}C, along with neat VE specimens tested at 30^{o}C, form a single cluster with an average tan delta value of 0.025. The remaining neat VE specimens tested at 60^{o}C, 90^{o}C, and 120^{o}C form the fourth cluster. In Figure

4.8b, a “scale data and display image (imagesc) object” plot is presented to indicate the number of clusters (each distinct set of bands in a row) and the bands associated with each cluster. The bands reflect the densities of data points within each cluster and correspond to the distances between the data points in Figure 4.8a. These findings prove that temperature is a dominant feature for the whole dataset.

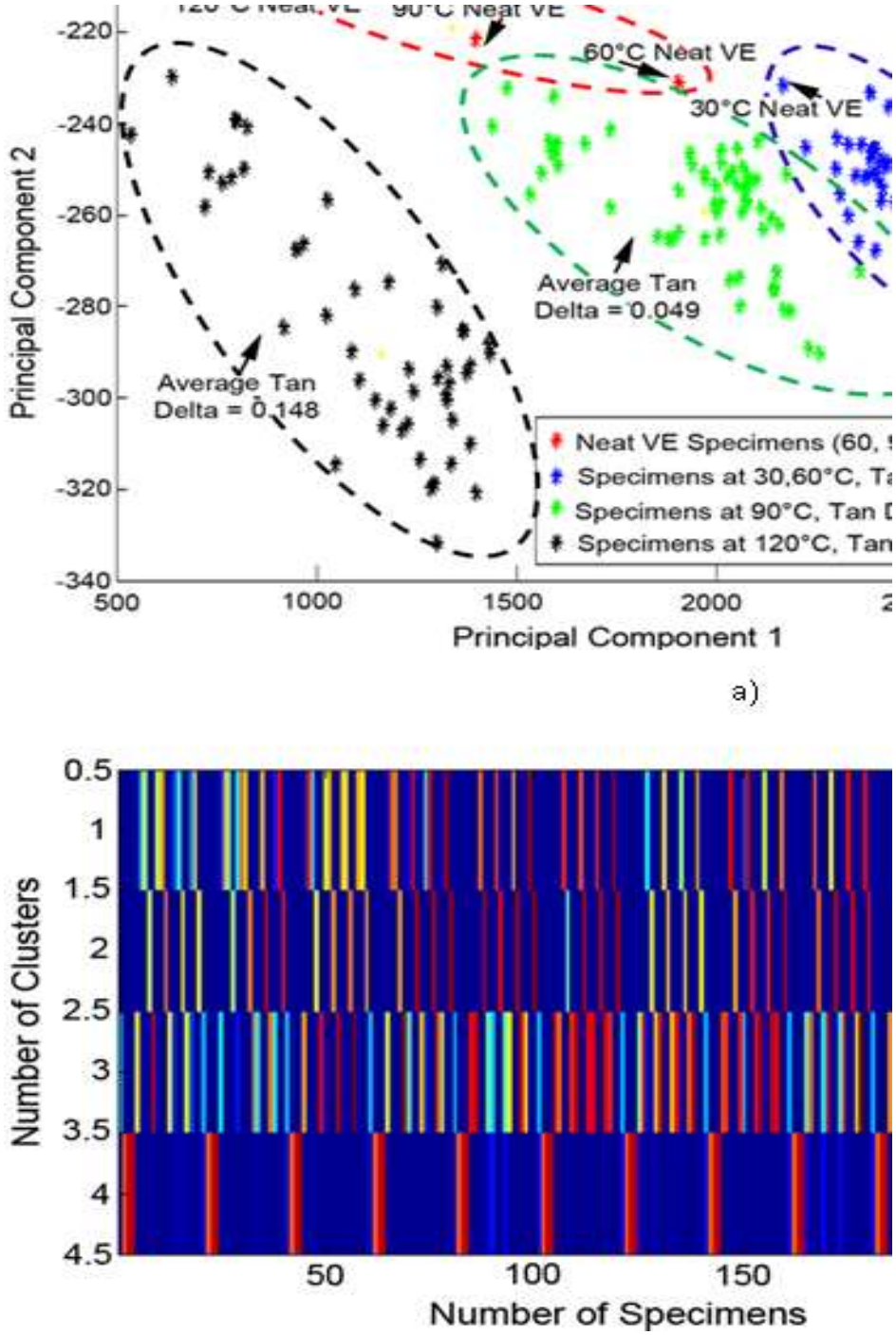
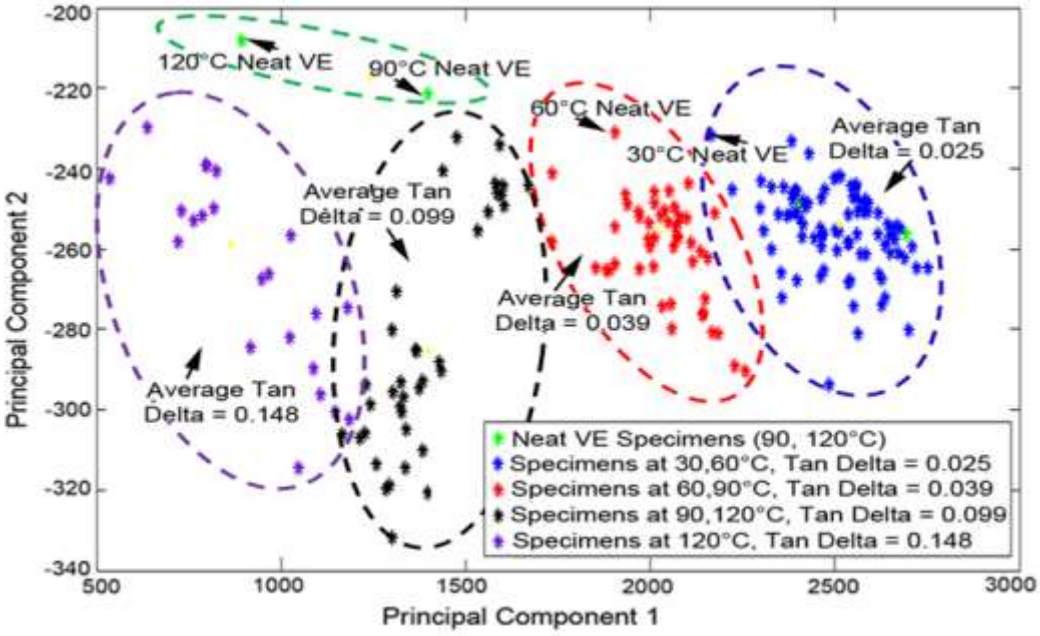


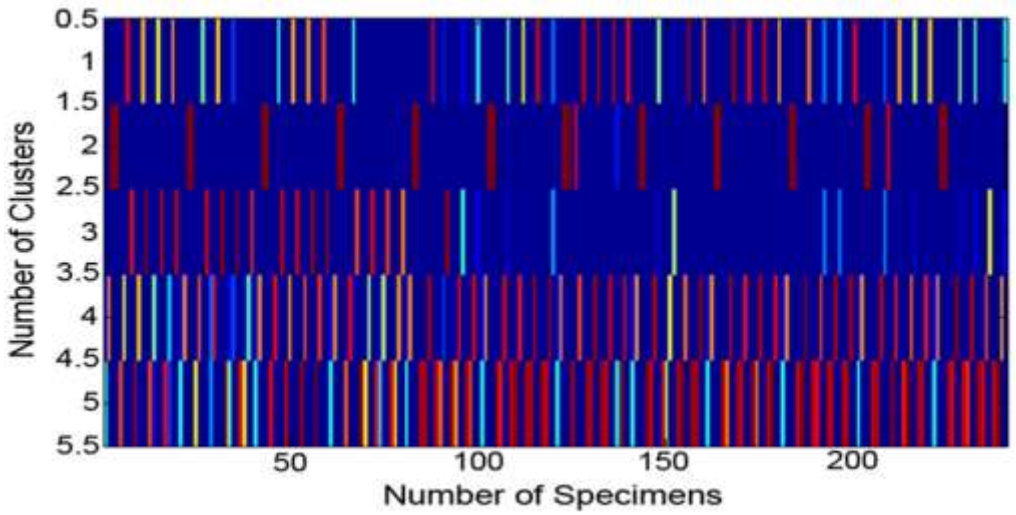
Figure 4.8 Clustering results and imagesc plot after applying the FCM algorithm and the GK distance measure, when $C = 4$.

Clustering results and temperature labels are included in a). b) is the "scale data and display image (imagesc) object" plot where four bands representing four clusters can be identified.

In Figure 4.9, the FCM results are illustrated where five clusters are chosen to represent the data using the GK distance measure. The nanocomposite and neat VE specimens tested at 30^{o}C form a cluster with an average tan delta value of 0.025. Included in this cluster is a fraction of the nanocomposite specimens tested at 60^{o}C. The remainder of the nanocomposite and neat VE specimens tested at 60^{o}C, along with a fraction of the nanocomposite specimens tested at 90^{o}C, are contained in a separate cluster with an average tan delta value of 0.039. The rest of nanocomposite specimens tested at 90^{o}C and a fraction of nanocomposite specimens tested at 120^{o}C form a third unique cluster with an average tan delta value of 0.099. The rest of the nanocomposite specimens tested at 120^{o}C form a fourth cluster with an average tan delta value of 0.148. Lastly, Figure 4.9a includes a fifth separate cluster that contains the neat VE specimens tested at 90^{o}C and 120^{o}C. In Figure 4.9b, an imagesc plot is presented, where five clusters can be identified. Again, these results demonstrate that temperature is a dominant feature.



a)



b)

Figure 4.9 Clustering results and imagesc plot after applying the FCM algorithm and the GK distance measure, when $C = 5$.

Clustering results and temperature labels are included in a). b) is the "scale data and display image (imagesc) object" plot where five bands representing five clusters can be identified.

Using the GK distance measure, FCM works better for the 240 VGCNF/VE specimens when the selected number of clusters equals four. For this case, specimens tested at different temperatures tend to be located in separate clusters that distinguish each of these temperatures. In addition, neat VE data specimens tested at 60-120^{o}C tended to cluster together. These results suggest that the FCM algorithm was able to identify VGCNF/VE specimens that have similar properties and placed them into different clusters.

In Figure 4.1, VGCNF/VE specimens tested at 60^{o}C and 90^{o}C are mixed up in the middle of the map. SOM analysis was run using only the specimens tested at 60^{o}C and 90^{o}C as shown in Figure 4.10.

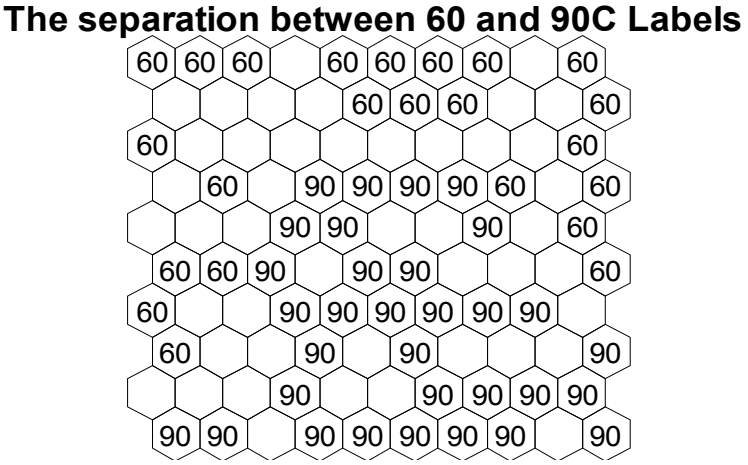


Figure 4.10 A 10 x10 SOM showing only VGCNF/VE specimens tested at 60 and 90^{o}C.

The separation between both groups can be clearly seen. Most of 60^{o}C specimens were clustered at the top of the map whereas 90^{o}C specimens were clustered at the bottom and at the middle of the map.

From Figure 4.10, a clear separation between the specimens tested at 60^{o}C and 90^{o}C can be seen. 60^{o}C specimens were clustered at the top and at both left and right

sides of the SOM whereas 90ºC specimens were clustered at the bottom and at the middle of the SOM. This reflects the fact that there is a physical distinction between both groups. In other words, when SOM was run using only VGCNF/VE tested at 60ºC and 90ºC, it was able to identify the specimens tested at both temperatures and hence it separated them in two groups and placed each specimen in a location in the map. The indices associated with each specimen in Figure 4.10 are illustrated in Figure 4.11. This will facilitate the comparison of physical and mechanical properties of the specimens that tend to cluster (group) together in the map (tested at the same temperature, 60ºC or 90ºC) as well as the common properties of the specimens that tend to form subclusters (subgroups) within each temperature group.

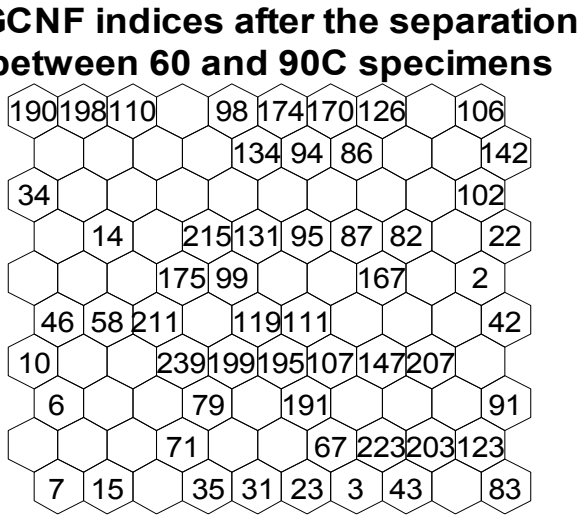


Figure 4.11 A 10 x10 SOM showing the corresponding indices of VGCNF/VE specimens tested at 60ºC and 90ºC in Figure 4.10

There are three specimens tested at 60ºC located at the top left corner of Figure 4.10 and their corresponding indices are 190, 198, 110 (Figure 4.11). These specimens

have the same tan delta value (0.02). In addition, in Figure 4.10, there are seven specimens tested at 90ºC located in the last two rows of the map. From Figure 4.11, their corresponding indices are 67, 223, 203, 103 (second row from the bottom of the map) and 43, 3, 23 (last row of the map). All these specimens have the same tan delta value (0.07). This means that, VGCNF/VE specimens, tested at either 60ºC or 90ºC, that tend to form subclusters within the main group share the same tan delta value. Tan delta values of the specimens tested at 60ºC and 90ºC are shown in Figure 4.12.

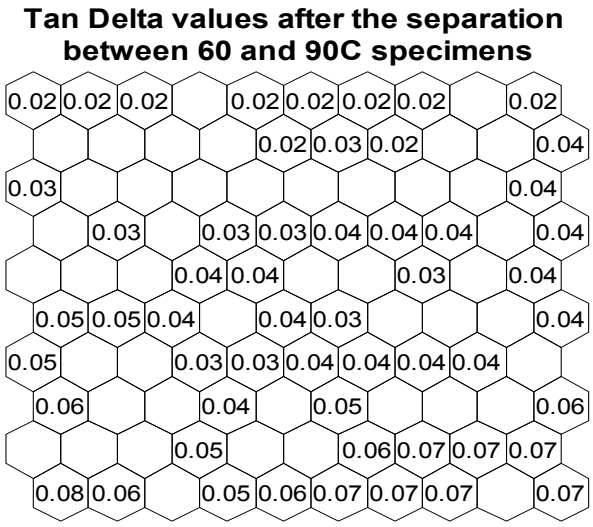


Figure 4.12 A 10 x10 SOM showing the corresponding tan delta values of the VGCNF/VE specimens tested at 60ºC and 90ºC in Figure 4.10.

From Figures 4.10 and 4.12, SOM was able to characterize the separation between both groups of specimens tested at 60ºC and 90ºC and within each group; specimens with the same tan delta value tend to form subclusters (subgroups). This means that temperature and tan delta are dominant features in this VGCNF/VE material

system. Thus, separating the specimens tested at these both temperatures has supported more this initial observation that was made earlier (see Figures 4.1, 4.3, and 4.4).

PCA analysis was run using the specimens tested at 60ºC and 90ºC (Figure 4.13).

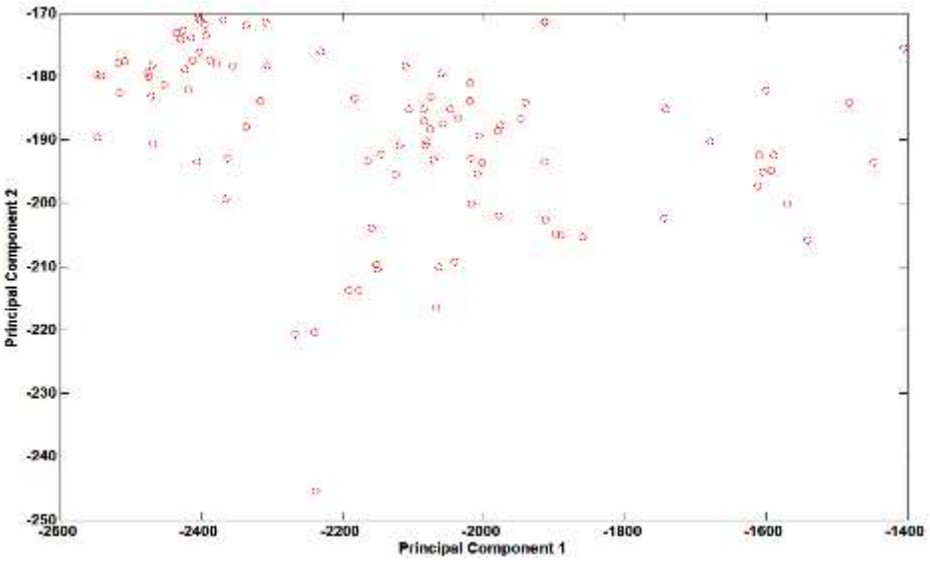
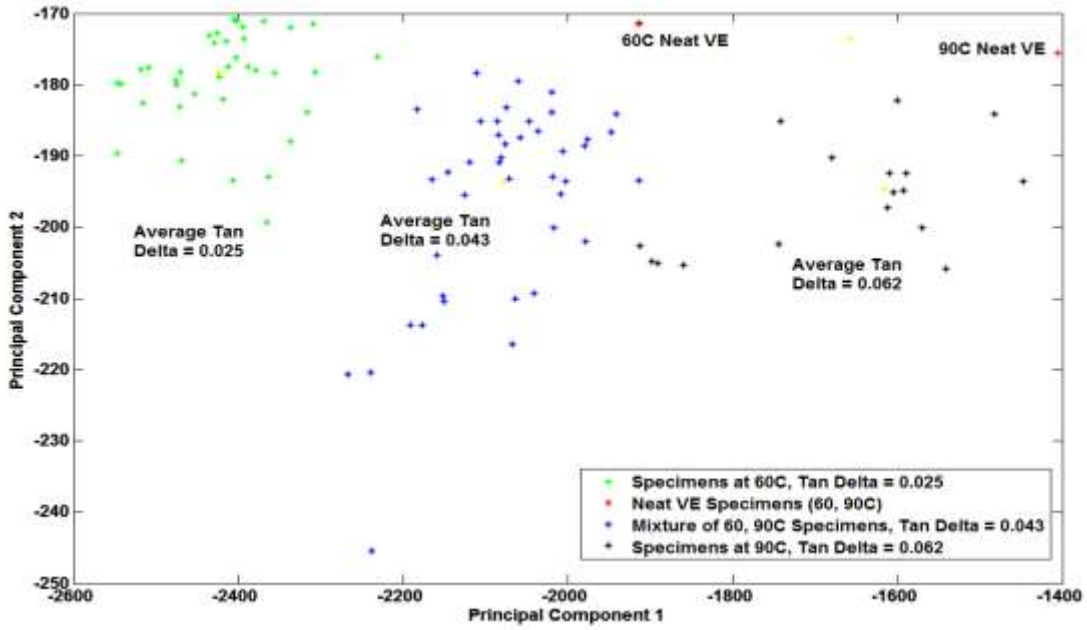


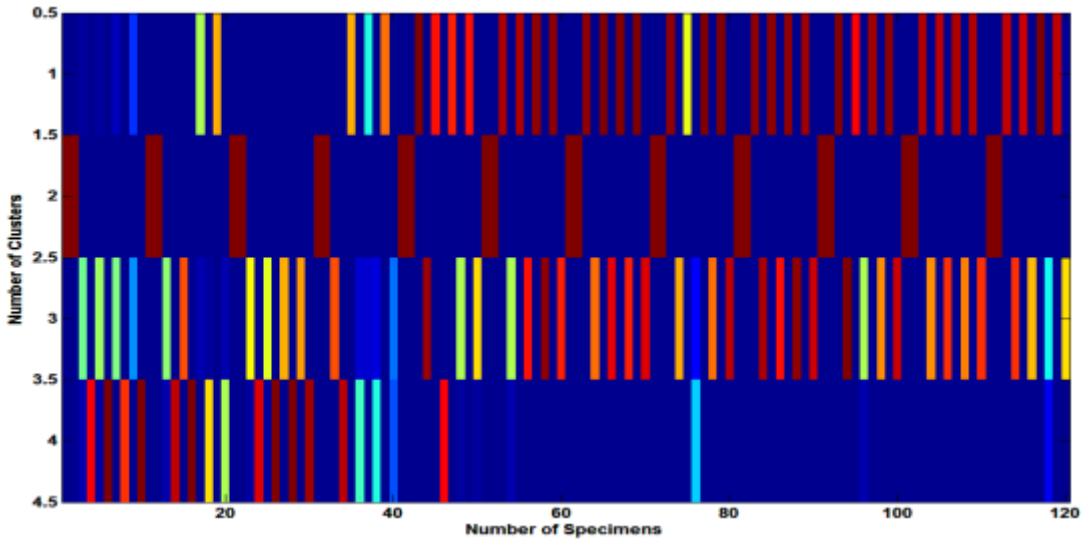
Figure 4.13 A 2-D graphical representation of the VGCNF/VE nanocomposite specimen data tested at 60ºC and 90ºC using the PCA technique

The FCM was applied to the VGCNF/VE nanocomposite data tested at 60ºC and 90ºC using the GK distance measures. In Figure 4.14, the FCM results are illustrated, where four clusters are chosen to represent the data tested at 60ºC and 90ºC using the GK distance measure. The data points are divided into four different clusters, each shown with a different color. In Figure 4.14a, the nanocomposite specimens tested at 60ºC and 90ºC each form a separate cluster on the left and right with average tan delta values of 0.025 and 0.062, respectively. The rest of the nanocomposite specimens tested at 60ºC and 90ºC are mixed and form a single cluster in the middle but with the same tan delta

value of about 0.043. Neat VE specimens tested at 60ºC and 90ºC form the fourth cluster. In Figure 4.14b, a "scale data and display image (imagesc) object" plot is presented to indicate the number of clusters (each distinct set of bands in a row) and the bands associated with each cluster.



a)



b)

Figure 4.14 Clustering results and imagesc plot after applying the FCM algorithm to the VGCNF/VE nanocomposite data tested only at 60°C and 90°C, when C = 4.

Clustering results and temperature and tan delta labels are included in a). b) is the "scale data and display image (imagesc) object" plot where four bands representing four clusters can be identified.

These findings give additional evidence that temperature is a dominant feature for the whole dataset and neat VE specimens tested at 60°C and 90°C have similar properties as they were all placed in one single cluster. In addition, tan delta is also a dominant feature because not only the clusters with specimens tested exclusively at the same temperature have the same tan delta value (0.025 and 0.062 for 60°C and 90°C clusters respectively), but also the cluster with a mix of specimens tested at both temperatures has the same tan delta value (0.043).

The SOM analysis allows a preliminary visual identification of the different existing groups [62]. In contrast, the FCM clustering approach identifies existing clusters and provides a mechanism to assign VGCNF/VE specimens to the appropriate cluster. Furthermore, FCM allows objects to belong to several clusters simultaneously, with different degrees of membership. This feature is not available in SOMs [64]. Hence, SOMs can be more helpful in identifying the dominant feature(s)/dimension(s) in the dataset. Other clustering algorithms (e.g., FCM) can be used to better identify cogent patterns and trends in VGCNF/VE data. In addition, different VGCNF/VE and/or neat VE specimens and their associated viscoelastic properties can be identified and categorized within their respective clusters. Each cluster can be identified based on one or more of the input design factors of the VGCNF/VE system.

Once again, for the ANN implementation in this research, the inputs and outputs were normalized using standardized scores, as their original value ranges were completely different from each other. This allowed the sigmoid functions to perform better. At the same time, the outputs were de-normalized. Because back-propagation is a gradient descent algorithm that can converge to a local optimum, the ANN model was

trained six times with different initial weights (set randomly), and the best results are reported here.

The overall ANN architecture is shown in Figure 4.15. This structure was used when both the resubstitution and the 3-folds CV techniques were implemented.

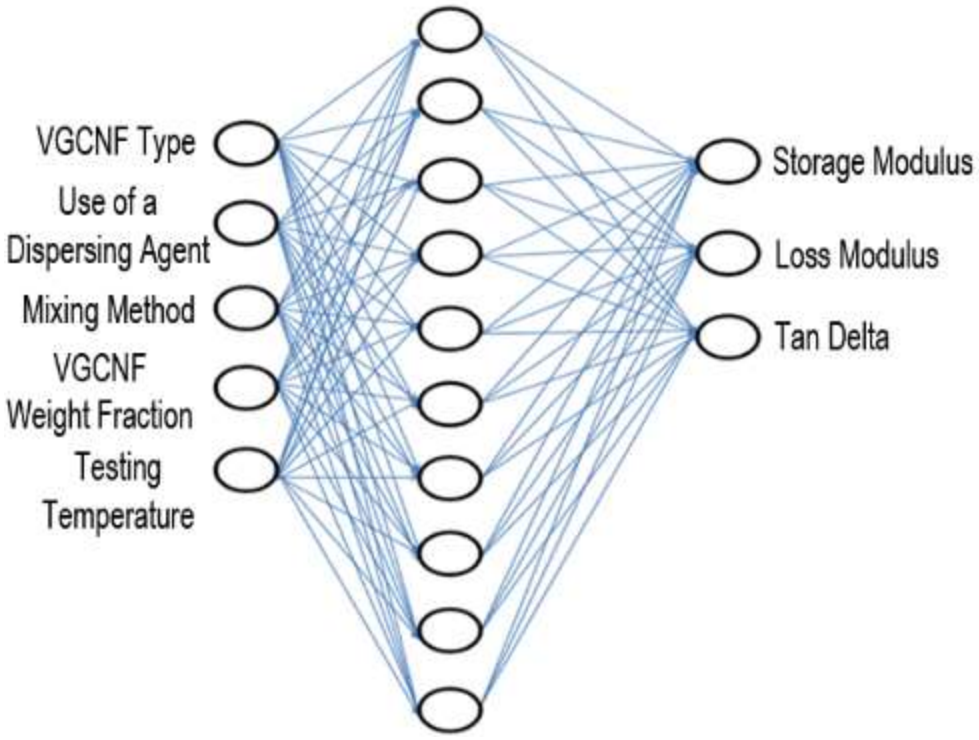


Figure 4.15 The architecture of ANN used in this study.

Five neurons were used in the input layer; each for one input and three neurons were used in the output layer; each for one response. There is one hidden layer with ten neurons.

There are three layers utilized in this ANN architecture (Figure 4.15). The input layer consists of five neurons and each neuron carries one of the inputs used in this study (VGCNF type, use of a dispersing agent, mixing method, VGCNF weight fraction, and testing temperature). There is one hidden layer consisting of ten neurons. The hidden layer connects the input layer with the output layer via activation functions from *input-*

hidden and from *hidden-output*. The output layer consists of three neurons, each for one of the responses used in this study (storage modulus, loss moduli, and tan delta).

First, the ANN was trained using the resubstitution method, where all the 240 VGCNF/VE samples were used for training and testing. The ANN implementation details are illustrated in Table 4.3.

Table 4.3 Implementation details of the BPANN applied to the VGCNF/VE dataset using the resubstitution method.

Number of input neurons	5
Number of output neurons	3
Number of hidden layers	1
Number of neurons in the hidden layer	10
Mean Square Error (MSE)	0.0015
Learning rate	0.001
Input-hidden activation function	Sigmoidal
Hidden-output activation function	Sigmoidal
Number of epochs	23

Each input in the VGCNF/VE dataset (VGCNF type, use of a dispersing agent, mixing method, VGCNF weight fraction and testing temperature) is associated with one input neuron in the input layer and each response (storage modulus, loss modulus, and tan delta) is associated with one output neuron in the output layer. Thus, the ANN architecture has five input neurons and three output neurons. The performance curve of the ANN implementation using the resubstitution method is shown in Figure 4.16.

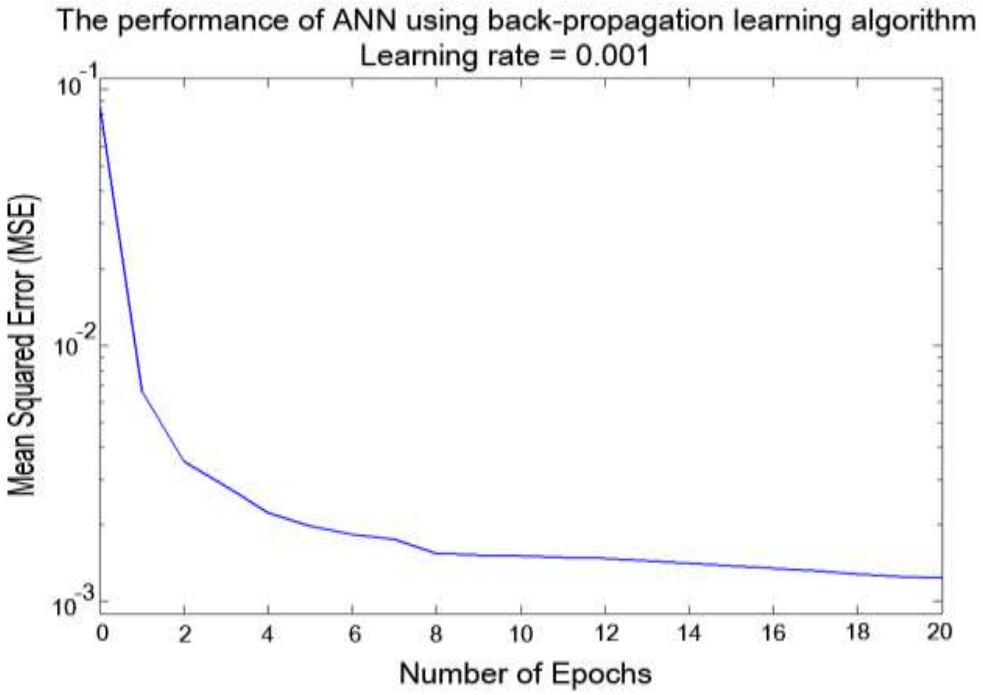


Figure 4.16 The performance curve of back-propagation ANN (BPANN) using the resubstitution method.

The MSE gradually decreases with the increasing number of epochs and the best performance (lowest MSE) is achieved after running 23 epochs.

After the resubstitution method, the 3-folds CV technique was applied on the VGCNF/VE dataset. Since the total number of samples was 240 and three different trials were implemented, the size of the training and test sets in each trial were 160 and 80 samples, respectively. Each trial had different training and test samples than those of the other trials. This led to a more efficient training and testing of the ANN model and, therefore, the best performance and network structure was obtained.

Similar to the ANN implementation using the resubstitution method, in 3-folds cross validation implementation, five neurons were used in the input layer (each VGCNF/VE input with one neuron) and three neurons in the output layer (each VGCNF/VE response with one neuron). The learning rate was 0.001 and the activation

function implemented between the input-hidden and hidden-output layer was the sigmoidal function. There was one hidden layer with ten neurons.

The performance curve of the ANN implementation using the first fold of the 3-folds cross validation technique is shown in Figure 4.17.

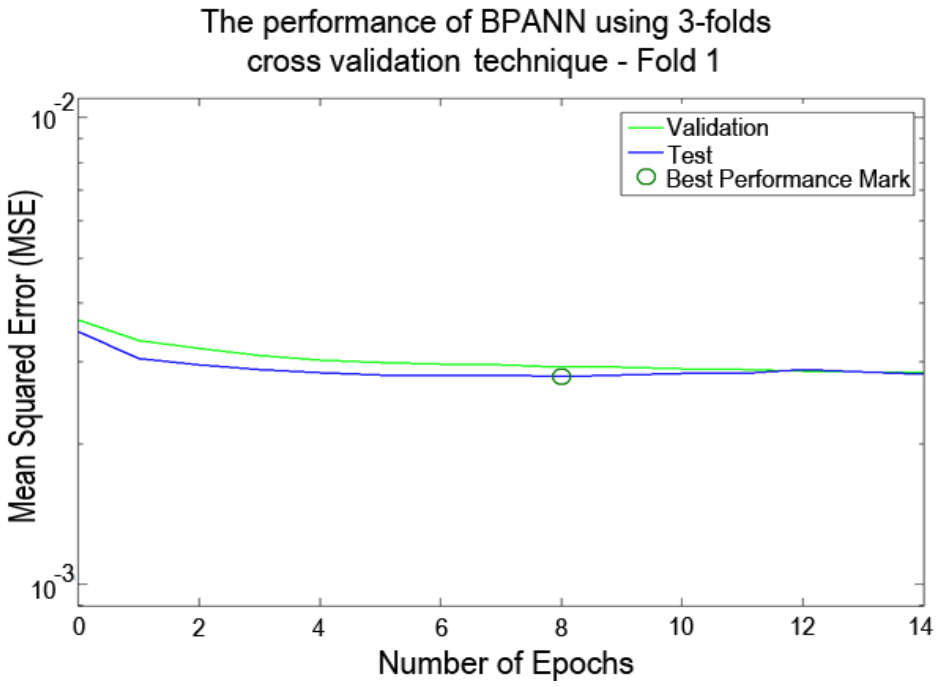


Figure 4.17 The performance curve using the data samples of the first fold when 3-folds cross validation technique was applied.

The performance of the training set is slightly better than that of the test set and the best performance was achieved at epoch 8 when the MSE reached about 0.0028 mark.

As evident in Figure 4.17, the performance of both test (validation) and training sets were very close, even though the training set behaved better up to epoch 10. After that, the MSE of the training set began to increase and was equal to the MSE of the test set at epochs 12-14. The best performance of the ANN model using the first fold was achieved at epoch 8, when the MSE was minimal at about 0.0028.

The performance curve of the ANN implementation using the second fold of the 3-folds cross validation technique is shown in Figure 4.18.

The performance curves were almost steady for both training and test sets (Figure 4.18). However, the training set behaved better than the test set compared to the training and test sets in the first fold. The best performance of the ANN model using the second fold was achieved at epoch 17, when the MSE was minimal at about 0.0029. After that, the MSE tended to slightly increase and remained constant up to epoch 23.

The performance curve of the ANN implementation using the third fold of the 3-folds cross validation technique is shown in Figure 4.19. In this figure, the total number of epochs needed for the ANN model to converge was 25 epochs and the MSE was minimal at epoch 21. The performance of both the training and test sets was nearly the same at the beginning, but later it was slightly better for the training set than for the test set. However, both curves remained almost steady after running seven epochs of the analysis.

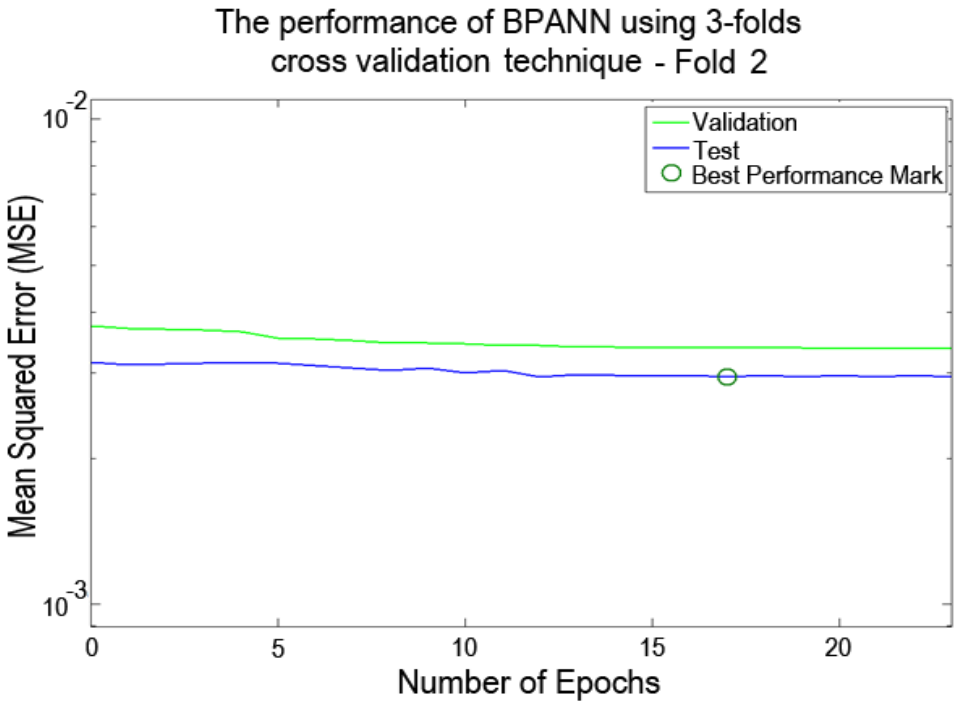


Figure 4.18 The performance curve using the data samples of the second fold when 3-folds cross validation technique was applied.

The performance of the training set is slightly better than that of the test set and the best performance was achieved at epoch 17 when the MSE reached about 0.0029 mark.

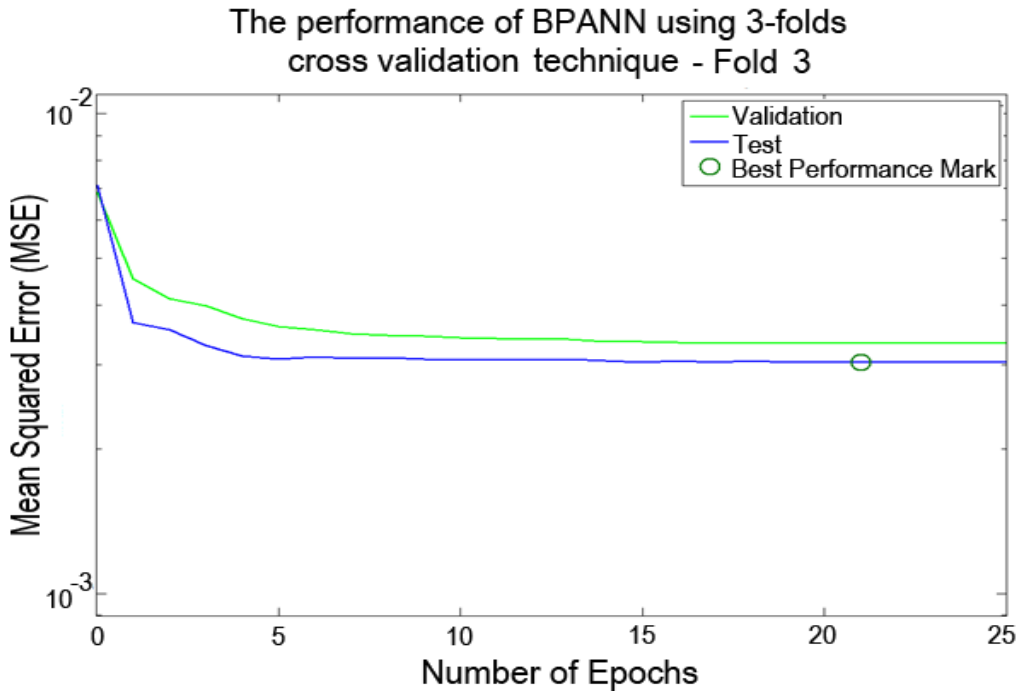


Figure 4.19 The performance curve using the data samples of the third fold when 3-folds cross validation technique was applied.

The performance of the training set is slightly better than that of the test set and the best performance was achieved at epoch 21 when the MSE reached about 0.0030 mark.

The number of epochs needed for the ANN model to converge was higher in the case where the resubstitution method was implemented. This is due to the fact that the number of training and test samples was higher than those of the 3-folds cross validation technique. However, this came at the expense of a lower MSE at 0.0015 using the resubstitution method, whereas the average MSE of all folds in the case where the 3-folds cross validation technique was implemented was higher at about 0.0029.

4.2 Advanced High Strength Steel (AHSS) Results and Analysis

As mentioned in Chapter 1, the initial attempt to utilize the concept of materials informatics and the underlying data mining and knowledge discovery techniques was to

develop 3GAHSSs using datasets from conventional, first, and second generations AHSS.

Conventional and first generation AHSS have a ferrite matrix with varying volume fractions of martensite or lower bainite as the higher strength reinforcement phase [76]. For transformation induced plasticity (TRIP) steels, small volume fractions of high carbon retained austenite which transforms to martensite during deformation. Examples of first generation AHSS include dual phase (DP), complex phase (CP), dual phase and martensitic (MART), and low alloyed TRIP steels [76].

Second generation AHSS have high Mn content (³ 18wt.%) and C contents (0.6wt%) and they frequently contain significant Si, Al and other alloying elements. Examples of second generation AHSS include twinning induced plasticity (TWIP), and lighter weight steels with induced plasticity (L-IP) steels [76].

Conventional AHSS are considered cost-effective steels but they do not sustain large quantities of strength-ductility balance (MPa%) over long periods of time despite the fact that some of them have a good elongation percentage. First generation AHSS sustain relatively higher amount of strength-ductility balance (≤ 20,000 MPa%) over time but they experience very low elongation percentage [76].

Second generation AHSS, however, has the best of both worlds. In this category, AHSS can sustain very high strength-ductility balance (50,000 – 80,000 MPa%) quantities over time along with very good elongation percentages compared to conventional and first generation AHSS. However, second generation AHSS are very costly and it is difficult to use them in large industrial scales within limited budgets [76].

An approximate illustration of strength-ductility balance along with elongation percentages for conventional AHSS and first and second generations AHSS is shown in Figure 4.20 [76]. The figure shows the trade-off between conventional and first generation AHSS on one side and second generation AHSS on the other side.

This work was focused in designing AHSS that can sustain high strength-ductility balance (MPa%) quantities over time along with high elongation percentages, but yet with reasonable costs. This new category of AHSS can be referred to as "Third Generation (3G) AHSS" and is shown in Figure 4.21.

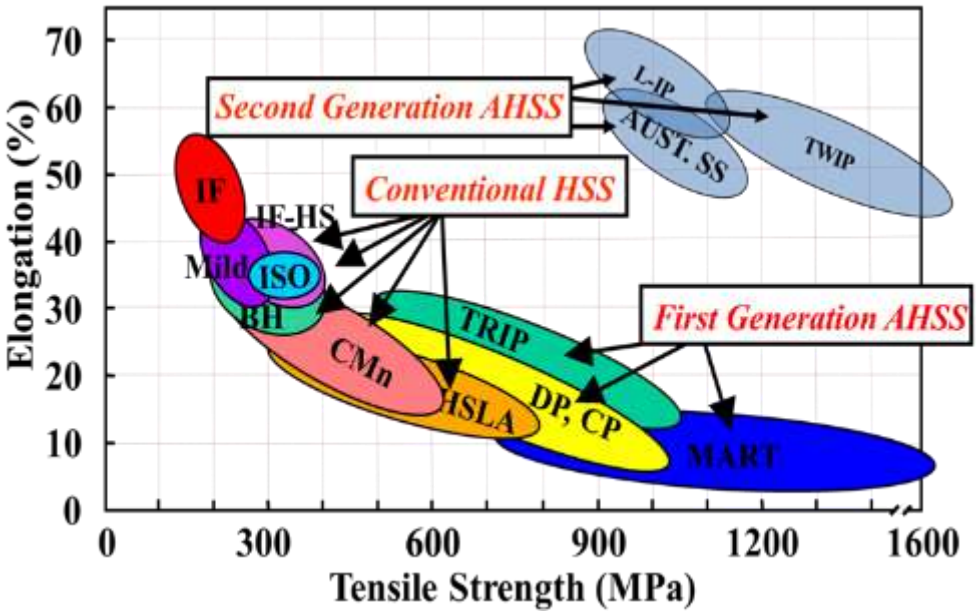


Figure 4.20 Current status of Advanced High Strength Steels (AHSS).
The chemical compositions and the generation for each AHSS are shown.

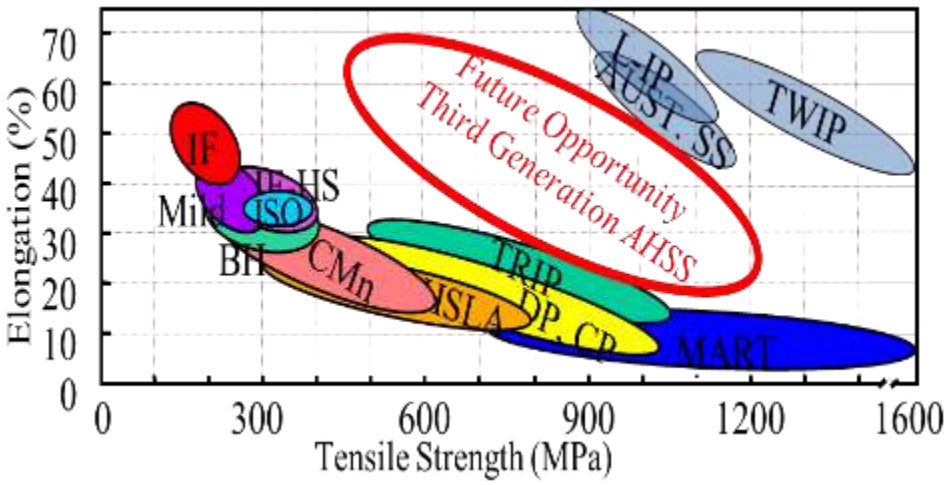


Figure 4.21 An illustration of the overall design goals of 3G AHSS

Less expensive and lower alloying elements than 2nd generation AHSS can be fabricated as well as higher mechanical properties than 1st generation AHSS can be achieved.

The dataset used in the study is shown in Table 4.4

Table 4.4 Conventional, first, and second generation AHSS dataset used in the initial study

Material			UTS (MPa)	Chemical Comp														
				(if left blank, paper did not include a value) also Fe is balanced														
				C	Mn	Si	P	S	Al	Nb	Cu	Cr	Mo	Ni	N	V	Total	Fe
TRIP 800	1		820	0.11	1.7	0.6	0	0	0	0	0	0	0	0	0	0	2.4259	97.5741
TRIP 600	2		650	0.08	1.7	1.6	0	0	0	0	0	0	0	0	0	0	3.374	96.626
TRIP	3		850	0.2	1.5	1.5	0	0	0	0	0	0	0	0	0	0	3.2	96.8
TRIP	4		920	0.2	1.5	1.5	0	0	0	0	0.5	0	0	0	0	0	3.7	96.3
TRIP	5		850	0.15	2.4	1	0	0	0	0	0	0	0	0	0	0	3.585	96.415
TRIP	6		800	0.14	2.2	0.5	0	0	0	0.018	0	0	0	0	0	0	2.895	97.105
TRIP	7		815	0.14	2.1	1	0	0	0	0.018	0	0	0	0	0	0	3.295	96.705
TRIP	8		625	0.21	1.8	0.3	0	0	1.2	0.003	0	0	0	0	0	0	3.504	96.496
TRIP 600	9		615	0.217	1.2	0	0	0	1.4	0	0	0	0	0	0	0	2.93	97.07
DP 600	10		600	0.072	1.6	0.2	0	0	0	0	0	0	0	0	0	0	1.955	98.045
DP	11		920	0.166	1	0.2	0	0	0	0	0	0	0	0	0	0	1.49	98.51
DP	12		775	0.065	1.6	0.4	0	0	0	0.026	0	0	0	0.6	0	0	2.664	97.336
DP	13		590	0.123	1.9	0.1	0	0	0	0	0	0	0.1	0	0	0	2.2316	97.7684
DP	14		597	0.16	1.4	2	0	0	0	0	0	0	0	0.1	0	0	3.587	96.413
DP	15		630	0.17	1.4	1.1	0	0	0	0	0	0	0	0	0	0	2.677	97.323
DP	16		645	0.16	1.4	1.7	0	0	0	0	0	0	0	0	0	0	3.295	96.705
DP	17		650	0.15	1.4	1.4	0	0	0	0	0	0	0	0.1	0	0	3.067	96.933
DP	18		627	0.15	1.4	1.2	0	0	0	0	0	0	0	0	0	0	2.824	97.176
DP	19		470	0.15	1.4	0.3	0	0	0	0	0	0	0	0.1	0	0	1.887	98.113
TWIP	20		880	0.3	25	0	0	0	0	0	0	12	0	0	0.4	0	37.7	62.3
MS	21		830	0.006	7	0	0	0	0	0	0	0	0	11	0	0	17.526	82.474
TWIP	22		1170	0.6	18	0	0	0	0	0	0	0	0	0	0	0	18.6	81.4
TWIP	23		940	0.6	18	0.2	0	0	1.3	0	0	0.4	0	0.1	0	0.11	20.61	79.39

The dataset in Table 4.4 consists of 19 data points (vectors) for 1st generation AHSS and 4 data points (vectors) for 2nd generation AHSS. The columns represent the features (dataset dimensions) and these features are the ultimate tensile strength (UTS) and the chemical compositions.

K-means clustering algorithm [69] and self-organizing maps (SOMs) [61, 62] were used in order to check if data points of similar physical, mechanical, or chemical properties were clustered (grouped) together after visualizing the dataset in two

dimensions via data dimensionality reduction techniques [63]. The clustering results are shown in Figures 4.22 and 4.23.

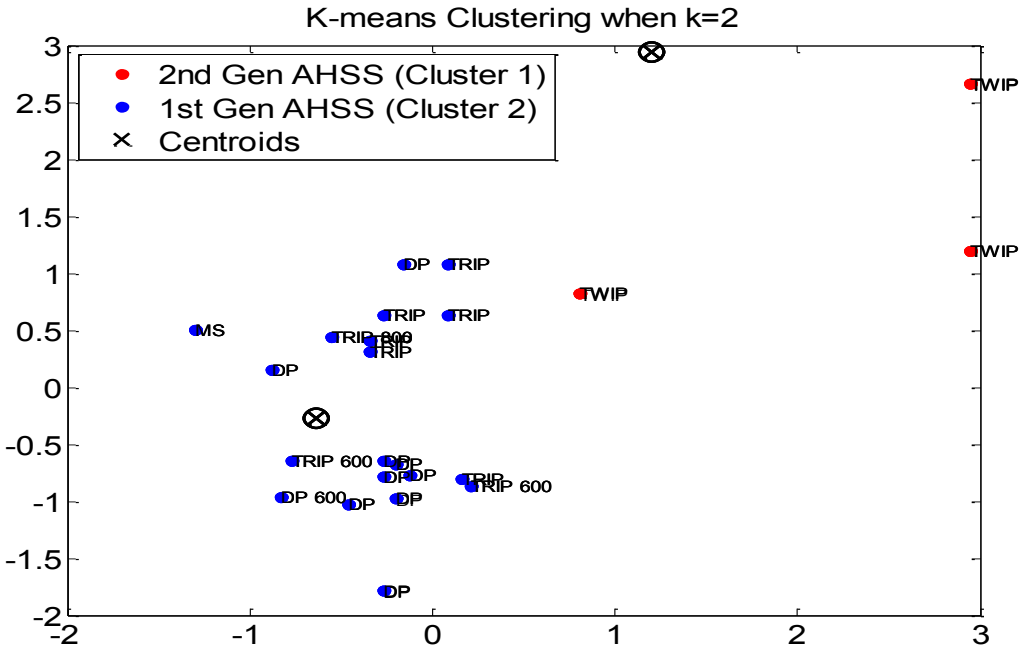


Figure 4.22 Clustering of AHSS dataset into two clusters (groups)

One cluster for conventional and first generation AHSS and one cluster for second generation AHSS.

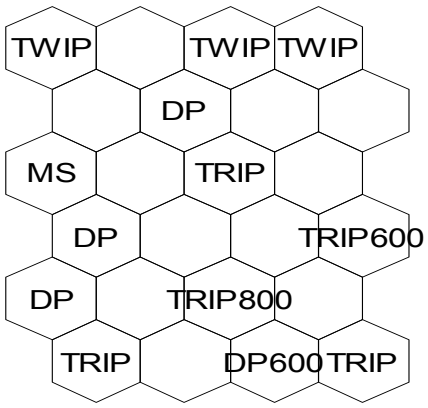


Figure 4.23 SOM implementation of AHSS dataset.

Samples with similar physical, mechanical, or chemical properties tend to be allocated close to each other in the map.

From Figure 4.22, data samples that belong to the same group were clustered together. In other words, conventional and first generation AHSS samples were clustered together in one group and second generation AHSS samples were clustered together in another group. This proves the ability of *K*-means clustering algorithm to cluster the samples of similar mechanical and physical properties together into separate clusters.

The SOM implementation in Figure 4.23 demonstrated that AHSS samples with similar properties were allocated together in the map. For example, TWIP samples were placed in the top raw of the map and TRIP samples were placed at the bottom of the map.

Furthermore, support vector machines (SVM) [70] technique was used in order to determine the classes to which AHSS samples in Table 3.2 belong. Resubstitution and 3-folds cross validation techniques were used in the analysis (Tables 4.5, 4.6, and 4.7).

Table 4.5 SVM classification performance when dot product kernel was implemented using different values of *C*.

Dot product kernel and C=0.5	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	100.00%	100.00%	100.00%	87.88%	0%
False Alarm Rate	0.00%	0.00%	0.00%	12.12%	100.00%
Dot product kernel and C=10	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	100.00%	100.00%	100.00%	87.88%	0%
False Alarm Rate	0.00%	0.00%	0.00%	12.12%	100.00%
Dot product kernel and C=1000	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	100.00%	100.00%	100.00%	87.88%	0%
False Alarm Rate	0.00%	0.00%	0.00%	12.12%	100.00%

Table 4.6 SVM classification performance when polynomial kernel of second degree was implemented using different values of *C*.

Polynomial kernel of degree 2 and C=0.5	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	100.00%	100.00%	100.00%	96.97%	0%
False Alarm Rate	0.00%	0.00%	0.00%	3.03%	100.00%
Polynomial kernel of degree 2 and C=10	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	100.00%	100.00%	100.00%	96.97%	0%
False Alarm Rate	0.00%	0.00%	0.00%	3.03%	100.00%
Polynomial kernel of degree 2 and C=1000	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	100.00%	100.00%	100.00%	96.97%	0%
False Alarm Rate	0.00%	0.00%	0.00%	3.03%	100.00%

Table 4.7 SVM classification performance when hyperbolic tangent kernel was implemented using different values of *C*.

Hyperbolic Tangent kernel and C=0.5	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	100.00%	75.00%	87.50%	81.82%	0%
False Alarm Rate	0.00%	25.00%	12.50%	18.18%	100.00%
Hyperbolic Tangent kernel and C=10	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	89.47%	100.00%	94.74%	81.82%	37.50%
False Alarm Rate	10.53%	0.00%	5.26%	18.18%	62.50%
Hyperbolic Tangent kernel and C=1000	**Resubstitution method**			**3-fold CV**	
	Class 1	Class 2	Average	Class 1	Class 2
Correct Classification Rate	89.47%	100.00%	94.74%	81.82%	37.50%
False Alarm Rate	10.53%	0.00%	5.26%	18.18%	62.50%

Using SVM technique, AHSS samples were classified into first and second generations AHSS with generally good classification rates using the resubstitution

method where all samples were used for training and testing. However, the classification rates were better when dot product and polynomial kernels were implemented than when the hyperbolic tangent kernel was implemented. The classification performance was not that good when the 3-fold cross validation technique was implemented because there were not enough data samples used for training and testing in each fold. Poor classification performance using the 3-fold cross validation reflects the fact that the number of AHSS samples is not enough to analyze the system and to establish a methodology to design/develop 3G AHSS.

There are, however, several problems that made this attempt unsuccessful to develop the 3G AHSS. First of all, there were not enough data samples (data points) of conventional, first and second generations of AHSS in literature and in experimental designs from which the desired properties of 3G AHSS can be inferred and utilized. Second, there was no correlation at all between the features used (i.e. between the chemical compositions and the ultimate tensile strength, UTS) as UTS relates more to microstructure *not* to chemical compositions except for elastic structures. Therefore, in order to design a framework for 3G AHSS, the features have to be modified and more samples have to be created (collected) which was not feasible at the time of this study.

CHAPTER V

APPLICATION TO MATERIALS INFORMATICS

The findings from the data mining confirm the trends established previously using surface response methodology [11, 12]. The dominant effect of temperature agrees with viscoelasticity theory in polymers [77], which states that the storage modulus drops steadily as temperature increases up to the glass transition temperature (T_g), where a sharp drop of several orders of magnitude occurs [78]. Concurrently, the loss modulus increases, reaching a maximum at T_g. The use of testing temperatures of 30, 60, 90, and 120 °C in this work allowed the significance of temperature to be clearly elucidated. The clustering of specimens tested at different temperatures proves the fact that at each temperature, the VGCNF/VE nanocomposite specimens tend to show similar viscoelastic behavior that is distinguishably different from their behavior at other temperatures. Of course, the effect of temperature is much larger than the effect of the other factors on the storage and loss moduli and hence, clustering based on temperature is more noticeable.

The effect of VGCNF weight fraction on both storage and loss moduli of VGCNF/VE nanocomposites were significant in previous studies [11, 12]. The storage modulus increases with increasing VGCNF weight fraction until a peak is reached at around 0.50 phr, the optimal VGCNF weight fraction. Since VGCNFs reinforce the polymer matrix, this stiffening of the matrix is expected. However, the presence of large VGCNF agglomerates at bigger fractions of VGCNF weight caused the storage modulus

values to not increase steadily [11]. The loss modulus typically decreases with increasing VGCNF weight fraction [11, 12]. However, due to the nature of the VGCNF agglomeration and dispersion in the polymer matrix and phenomena associated with it, such as stress concentration and frictional sliding in the entangled VGCNF networks [12], a more complex viscoelastic behavior is observed in VGCNF/VE nanocomposites. Nevertheless, the significance of VGCNF weight fraction is undisputable and was correctly discovered and matched the experimental trends. The clustering of neat VE specimens indicates once again that a sharp difference exists between the viscoelastic responses of these specimens with the VGCNF/VE nanocomposites as discussed above. Previously, a ~20% increase in the storage modulus had been observed by introducing VGCNFs into neat VE [12]. Furthermore, VGCNFs significantly modify the viscoelastic responses of neat VE at higher temperatures and VGCNF weight fractions, while the effect is not that pronounced for neat VE [12]. Hence, the neat VE specimens tend to cluster for testing temperatures of 60, 90, and 120 °C, matching previous observations [12].

The results and analyses of ANN implementation shown in Chapter V validate that all of the five input design factors and the three responses form a good combination to design a VGCNF/VE material system. This observation can be proved by the small value of MSE obtained through the analysis as well as the relatively few number of epochs (an average of 17.25 epochs) required in order for the ANN to converge (i.e. to achieve the minimal MSE). However, when tan delta response was excluded from the analysis, the MSE started to increase (about 0.02). In addition, when the testing temperature was removed as an input design factor, the MSE started to significantly

increase (about 0.24). At the same time, the number of epochs required for the ANN to converge was much higher (about 1000 epochs). On the other hand, when VGCNF weight fraction was excluded as a design factor, the average MSE increased, but not as much as when the testing temperature was excluded.

These observations confirm that the testing temperature is the most dominant feature in the input design factors followed by VGCNF weight fraction. Tan delta is also very important as a response. However, all other input design factors had exhibited less sensitivity, as their impacts on the responses were much less than the testing temperature and the VGCNF weight fraction. The findings in this study confirm previous findings by Nouranian, *et al*. [66], where a response surface modeling approach was utilized to optimize the VGCNF/VE nanocomposite material system.

Finally, in addition to demonstrating the dominant features in the data, the developed data mining models may be used to optimize the viscoelastic responses of the VGCNF/VE nanocomposites. It further guides the design and fabrication of VGCNF/VE nanocomposites. This work highlights the feasibility of knowledge discovery techniques in materials science and engineering and the rising field of materials informatics.

CHAPTER VI

THE PROPOSED DATA ANALYTICS METHODOLOGY AND OTHER NANOCOMPOSITES STRUCTURES

In this chapter, signal processing, data mining and knowledge discovery techniques were employed and extended to acquire new information about not only the viscoelastic VGCNF/VE (as has been studied in Chapters 3 and 4), but also about the flexural, and impact strengths properties for VGCNF/ VE nanocomposites. These properties were used to design a unified VGCNF/VE framework solely from data derived from a designed experimental study. Formulation and processing factors (curing environment, use or absence of dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time) and testing temperature were utilized as inputs and the true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta were selected as outputs.

6.1 Introduction

Earlier in the dissertation, data mining and knowledge discovery techniques were applied, as a proof of concept and as a first attempt to utilize the concept of materials informatics, to a VGCNF/VE nanocomposite system [10, 11, 66, 13] where formulation and processing factors (VGCNF type, use of a dispersing agent, mixing method, and

VGCNF weight fraction) and testing temperature were utilized as inputs and the storage modulus, loss modulus, and tan delta were selected as outputs.

In this chapter, the viscoelastic nanocomposite dataset was expanded into a unified framework which includes more VGCNF/VE structures and then data mining and knowledge discovery techniques were applied to the resulting dataset. The new expanded framework consists of the viscoelastic VGCNF/VE data, impact strengths data of VGCNF/VE [13], and flexural properties of VGCNF/VE [79]. This is the first time that such framework is designed, studied and analyzed and the major contribution of this paper is to apply data mining and knowledge discovery techniques in order to discover new knowledge, properties, and trends that have not been known *a priori* using this framework, thereby aiding the nanocomposite design, fabrication, and characterization without the need to conduct expensive and time-consuming experiments.

In this new study, several signal processing and supervised and unsupervised knowledge discovery techniques were used to explore an expended VGCNF/VE framework [10, 13, 79]. The dataset in the framework consisted of 565 data points each corresponding to the combinations of eight input design factors and nine output responses, i.e., a total of seventeen "dimensions." The dimensions in data mining are the combination of both inputs and outputs of the developed model. The dimensions of the new VGCNF/VE framework are curing environment, use or absence of dispersing agent, mixing method, VGCNF fiber loading (weight fraction), VGCNF type, sonication time, temperature, high-shear mixing time, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta (ratio of loss to storage modulus),

where the last nine dimensions correspond to measured macroscale material responses. Kohonen maps [61, 62], or self-organizing maps (SOMs), were reapplied to the new dataset in order to conduct a sensitivity analysis of all of these factors and responses. In addition, principal component analysis (PCA) [63] was reused to provide a two-dimensional (2-D) representation of nanocomposite data. This facilitated application of the fuzzy C-means (FCM) clustering algorithm [64, 65] to characterize the physical/mechanical properties of the new VGCNF/VE nanocomposites framework.

6.2 Materials and Methods

A brief summary of the statistical experimental design and testing procedures to generate the newly designed VGCNF/VE framework is given here. A more detailed discussion can be found in [10, 11, 66, 13, 79].

6.2.1 Statistical Experimental Design

The newly designed VGCNF/VE framework consists of VGCNF/VE viscoelastic data, VGCNF/VE flexural data, and VGCNF/VE compression and tension impact strengths data.

In addition to the 240 viscoelastic VGCNF/VE treatment combinations studied earlier in this dissertation, a total of 172 compression impact strengths, 93 tension impact strengths, and 60 flexural data treatment combinations were added to the newly designed framework. Therefore, the new VGCNF/VE framework has a total of 240+172+93+60= 565 treatment combinations [13, 79]. Different data interpolation techniques [80] were used to replace some of the missing and unknown data fields in the new framework.

6.2.2 Materials and Processing

For the VGCNF/VE viscoelastic specimens, a low styrene content (33 wt%) VE resin (Ashland Co., Derakane 441-400) and two VGCNF commercial grades. Namely, pristine PR-24-XT-LHT and surface-oxidized 24-XT-LHT-OX (Applied Sciences Inc.). These grades were utilized for nanocomposite specimen preparation [11, 66]. In addition, methyl ethyl ketone peroxide (MEKP) (US Composites Inc.) and 6% cobalt naphthenate (CoNaph) (North American Composites Co.) were selected as initiator and crosslinking promoter, respectively. Air release additives BYK-A 515 and BYK-A 555 (BYK Chemie GmbH) were used to remove air bubbles introduced during mixing. A commercial dispersing agent BYK-9076 (BYK-Chemie GmbH) was employed to improve VGCNF dispersion in the resin.

From a group of resin comprising 100 parts resin, 0.20 phr 6% CoNaph, 0.20 phr BYK-A 515, 0.20 phr BYK-A 555, 0.00-1.00 phr VGCNFs (based on the design given in Table 3.1), and a 1:1 ratio of BYK-9076 to VGCNFs, specimens used for testing were fabricated. The VGCNF/resin blend was mixed by either an ultrasonicator whose model is GEX750-5C from Geneq Incorporation, high shear mixer whose model is L4RT-A from Silverson Machines Ltd., or a combination of both, as dictated by the design given in Table 3.1. Then the nanofiber/resin blend was degassed under vacuum for 5-15 min at pressures of 8-10 kPa. The blend was thermally cured for 5 h at 60˚C followed by 2 h post-curing at 120˚C.

For the compression and tension impact strengths specimens, the VE resin chosen is 441-400 Derakane which contains 33 wt % styrene and it is produced from Ashland Incorporation in Covington, KY. Using VGCNFs (Model: PR-24-XTLHT-OX from

Applied Sciences Incorporation in Cedarville, OH) that were processed by surface-oxidized high temperature, this resin was reformed. Furthermore, the mixture had two air release agents whose models are BYK-A 515 and BYKA 555 from BYK USA, Incorporation in Wallingford, CT, a catalyst promoter which contains 6 wt % cobalt naphthenate processed in styrene solution and is produced from North American Composites in Lino Lakes, MN, a polymerization initiator which is methyl ethyl ketone peroxide (MEKP) from U.S. Composites, Incorporation in West Palm Beach, FL, and a dispersing agent whose model is BYK-9076, from BYK USA, Incorporation. These elements were mixed into a 125 g of resin [13].

Impact specimens were manufactured from a 125 g batch of resin. The VGCNFs and a dispersing agent were added as indicated by the design procedure and initially mixed with the resin manually. Then, VGCNF/resin batches were blended using a high shear mixer whose model is L4RT-A from Silverson Machines Incorporation in East Longmeadow, MA at 4500 rpm, an ultrasonicator whose model is GEX750-5C from GENEQ Incorporation in Montreal, Canada operating continuously at 20% of power amplitude, or a combination of both. In Design procedures that utilized both mixing methods, ultrasonication process was performed after the high shear mixing process. During the period of high shear mixing, resin mixture was placed in an ice bath. This will significantly eliminate heating the resin during processing phase. Ice bath absorption was also used for ultrasonication process 8.79 minutes or longer. The last step involved adding and manually mixing the MEKP initiator for few minutes. The whole mixture was then vacuumed and degassed until all air bubbles disappeared completely from the surface. Finally, the mixture was poured into the mold. To cure the mixture, it was placed

in a preheated oven which is produced by Fisher Scientific in Pittsburgh, PA. The nanocomposite specimens were placed in a nitrogen atmosphere for 5 hours at the temperature of 60°C for curing and then for 2 hours at the temperature of 120°C for postcuring [13].

For the flexural specimens, because of its excellent thermal characteristics and its resistance to corrosion, a VE infusion resin (from Ashland Chemical 441-400, Derakane) whose molecular weight is about 690 g/mol and 33% styrene was chosen for the matrix formulation. Pristine PR-24-XT-LHT and surface-oxidized 24-XT-LHT-OX were both bought from Applied Sciences, Inc. and were not modified. The specifications of XT-24-LHT-PR are: a surface area of about 35–45 m^2/g, a diameter of about 150 nm, and a dispersive energy of about 155 mJ/m^2. These quantities were specified in the manufacturer's manual. The manufacturer has not included the corresponding data for the oxidized 24-XT-LHT-OX. Despite the exclusive nature of the functional components exist on the oxidized nanofibers surface, lactones, carboxylic acids, and ring oxygen ethers, phenolic hydroxyl groups, quinones, and ketones may be included as well. By interacting with the oxygen-containing polar groups within the VE molecules structures, these functional components may improve nanofiber matrix linkage. Because VGCNF mechanical properties are highly resistant to oxidative treatment operation, only changes in the nanofiber to matrix linkage can cause the nanofiber oxidation on the properties of composite structures. A 6 wt% cobalt naphthenate solution in styrene, produced from North American Composites, was used as the promoter and methyl ethyl ketone peroxide (MEKP), produced from US Composites, was used as a free radical initiator to cure the resin. In order to minimize void formation in the final cured nanocomposite structure, the

air release additives BYK-A 515 and BYK-A 555 (from BYK Chemie GmbH), were used. BYK-9076 (from BYK Chemie GmbH) is an exclusive copolymer alkylammonium salt utilized as an active agent and has been used for different nanocomposite formulations. Thus, BYK-9076 was used as a dispersing agent (DA) for VGCNFs. Based on the directions set by the manufacturer, the ratio of 1:1 is the optimal ratio of the amount of DA to that of VGCNFs [79].

Another note is that from a single group of VGCNF/VE mixture of each treatment combination structure shown in Table 3.1, the flexural specimens were fabricated.

6.3 Theory/ Calculation

This new study incorporates eight input design factors, i.e., curing environment (nitrogen, oxygen), use or absence of dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time, and testing temperature and nine output responses, i.e., true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta. Hence, the dataset represents a seventeen-dimensional (17-D) space for analysis. Since curing environment, use or absence of dispersing agent, mixing method, and VGCNF type are considered qualitative factors, they are represented by a numeric code for analysis purposes. For two-level factors (curing environment, use or absence of dispersing agent, and VGCNF type), 0 and 1 are the coded values for the first and second levels, respectively. For the three-level factor (mixing method), -1, 0, and 1 are the coded values for the first, second, and third levels, respectively (Table 3.1).

On the basis of the above discussion, SOMs [61, 62], PCA [63], and the FCM clustering algorithm [64, 65] were used with the 565 treatment combination dataset to discover data patterns and trends for the expanded nanocomposites framework and to identify the different system features related to the specific material properties. SOMs were created with respect to VGCNF fiber loading, high shear mixing time, sonication time, temperature, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta. After analyzing the SOMs, temperature was identified as the most important input feature for the VGCNF/VE nanocomposites new framework because it has the highest impact on the resulting responses. VGCNF high shear mixing time and sonication time were also important features. In addition, it was inferred from the SOMs that some specimens tested at the same temperature tended to have several sub-clusters (groups). Each sub-cluster had the same tan delta or high shear mixing time or sonication time values. In addition, after analyzing the clustering results, it has been found that the viscoelastic VGCNF/VE data is very important in the newly designed VGCNF/VE framework.

6.4 Results and Discussions

In Figure 6.1, a 10x10 SOM resulting from the 565 data points is shown. Nanocomposite specimens tested at the same DMA temperature tend to cluster together. For example, specimens tested at 30^{o}C tend to cluster at the top, at the middle, and at the lower left corner of the map, whereas specimens tested at 90^{o}C and 120^{o}C tend to cluster at the lower right corner. Most importantly, since the SOM contains many specimens which were tested at 30^{o}C, this gives an indication that 30^{o}C is the temperature that has

the highest impact on the characteristics and properties of the studied nanocomposites specimens in the designed framework. The testing temperature of 120°C is also important (since the SOM has a small cluster of 120°C in the lower right corner) but not as critical as 30°C.

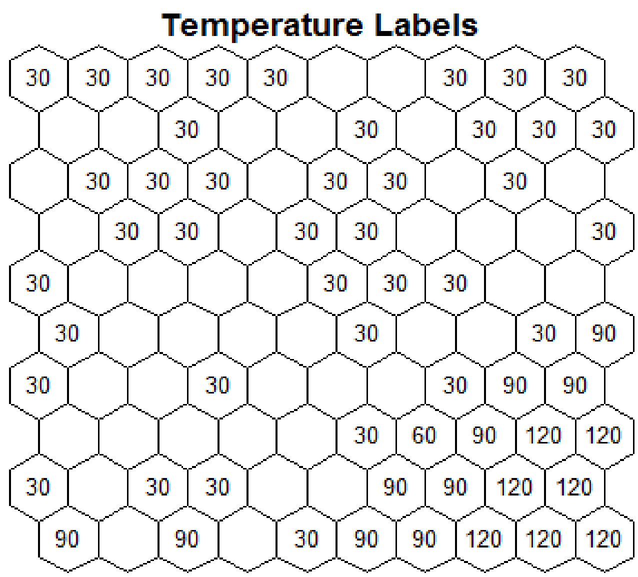


Figure 6.1 A 10×10 SOM with respect to temperature

This is for the 565 nanocomposite specimens used in the study (with all seventeen dimensions). The specimens tested at the same temperature tend to cluster together and 30°C is the testing temperature that drives the characteristics of all specimens in the designed framework.

In Figures 6.2, 6.3, and 6.4 three 10×10 SOMs for the VGCNF high shear mixing time, VGCNF sonication time, and the tan delta response are shown, respectively. In Figure 6.2, specimens with the same or close values of high shear mixing time tend to cluster together especially at the top and at the middle of the SOM. This means that high shear mixing time is important in the newly designed VGCNF/VE framework. However, this tendency is not consistent and is less than the clustering tendency shown in Figure

6.1 for temperature. Similarly, in Figure 6.3, specimens with the same or close values of sonication time tend to cluster together. However, compared with Figure 6.2, the clustering tendency of the specimens based on the sonication time is more pronounced than that of the high shear mixing time but less than the clustering tendency for temperature in Figure 6.1. In Figure 6.4, VGCNF/VE specimens with the same tan delta response values tend to cluster together and the clustering tendency is more consistent than that of high shear mixing time and sonication time. This leads to the conclusion that both tan delta and temperature are dominant features for the treatment combinations and have the highest impact on the responses for all the specimens in the framework followed by the sonication time and then by the high shear mixing time.

Another observation that can be seen from Figure 6.4 is that most specimens in the SOM have a tan delta of 0.00 which were clustered at the top and at the middle of the map. This means that the specimens with no delta values, i.e. the impact strengths specimens and the flexural specimens are essential components in the new framework. In addition, by comparing Figure 6.1 with Figure 6.4, most specimens of testing temperature 30^{o}C have a corresponding 0.00 tan delta response value. This leads to the conclusion that impact strength specimens and flexural specimens treated at 30^{o}C are very important in the new nanocomposites framework.

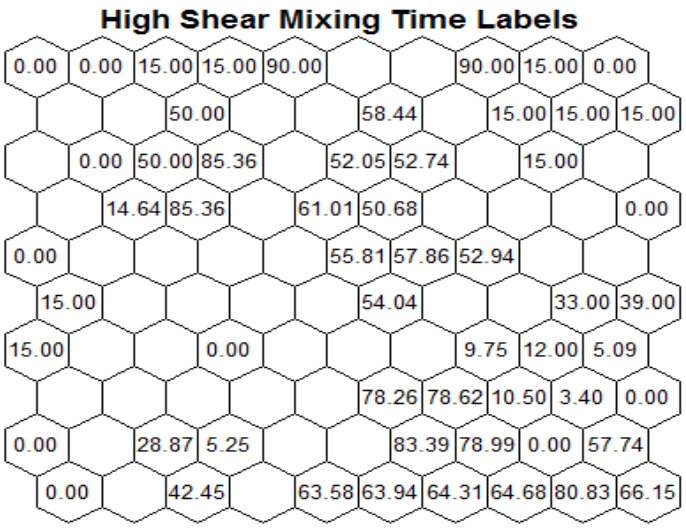


Figure 6.2 A 10×10 SOM with respect to VGCNF high shear mixing time.

The clustering tendency is less than that of the temperature in Figure 6.1 and can be seen clearly at the top and at the middle of the SOM.

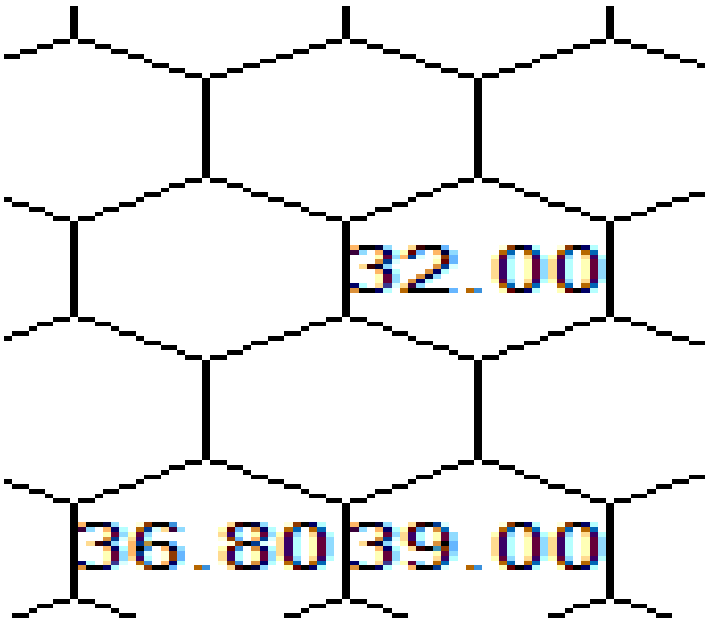


Figure 6.3 A 10×10 SOM with respect to VGCNF sonication time.

The clustering tendency is less than that of the temperature in Figure 6.1 but more than that of high shear mixing time in Figure 6.2 This can be clearly seen at the top, and the middle, and at the lower right corner of the map.

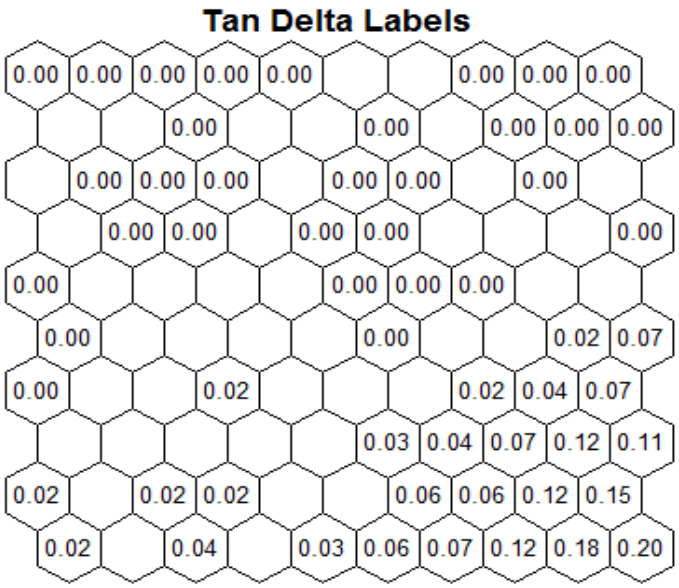


Figure 6.4 A 10×10 SOM with respect to tan delta values.

The clustering tendency is similar to that of the temperature in Figure 6.1. However, specimens treated at 30°C have a corresponding 0.00 tan delta value. This means that impact strengths and flexural specimens tested at 30°C are essential components in the new nanocomposites framework.

In Figure 6.5, a 10×10 SOM for the VGCNF fiber loading (sometimes referred to as VGCNF weight fraction) is shown. Unlike the sound impact of VGCNF weight fraction on the viscoelastic properties of VCCNF/VE based on the study conducted by AbuOmar *et. al* [81], this feature does not have that much impact on the new nanocomposites framework as the clustering tendency is inconsistent throughout the SOM.

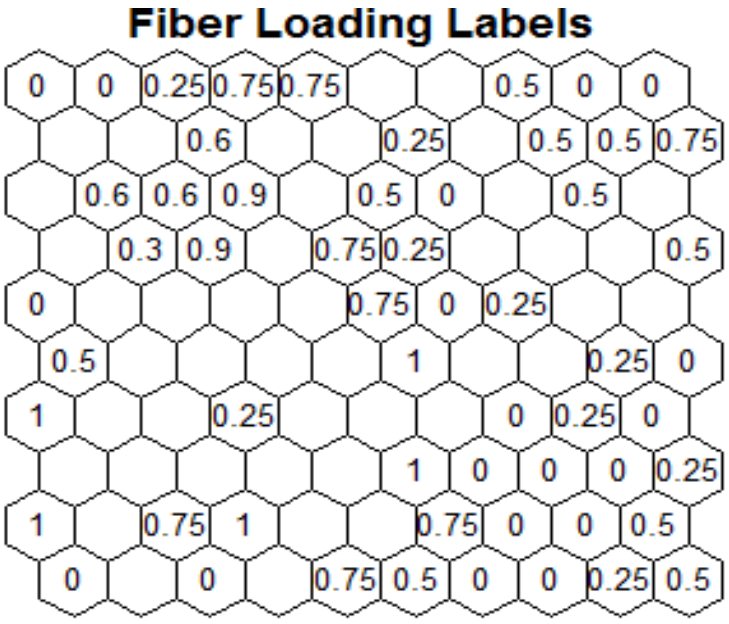


Figure 6.5 A 10×10 SOM with respect to VGCNF fiber loading (VGCNF weight fractions) values.

The clustering tendency is inconsistent throughout the SOM and so VGCNF fiber loading is *not* dominant for the treatment combinations of the newly designed nanocomposites framework.

In Figures 6.6, 6.7, 6.8, 6.9, 6.10, 6.11, 6.12, and 6.13, eight 10×10 SOMs for the responses of true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, and loss modulus are shown, respectively. From Figures 6.6, 6.7, 6.8, and 6.9 specimens that have no true ultimate strength responses, no true yield strength responses, no engineering elastic modulus responses, and no engineering ultimate strength responses are important in the new framework. This can be seen in four large clusters of 0.00 in Figures 6.6, 6.7, 6.8, and 6.9. This means that VGCNF/VE flexural and viscoelastic specimens are essential in the new framework. Similarly, from Figures 6.10 and 6.11, one can see that specimens with no flexural modulus and flexural strength responses are also important in

the new framework. Particularly, impact strengths specimens and viscoelastic specimens are essential in the new design. In addition, from Figures 6.12 and 6.13, specimens that don't have any responses for storage modulus and loss modulus are important in the whole framework; namely the impact strengths specimens and the flexural specimens. This leads to the conclusion that all the components (i.e., VGCNF/VE impact strengths, flexural, and viscoelastic specimens) of the newly designed nanocomposites framework are essential. Therefore, the validity of the selection of these nanocomposites structures in the framework is confirmed by this observation.

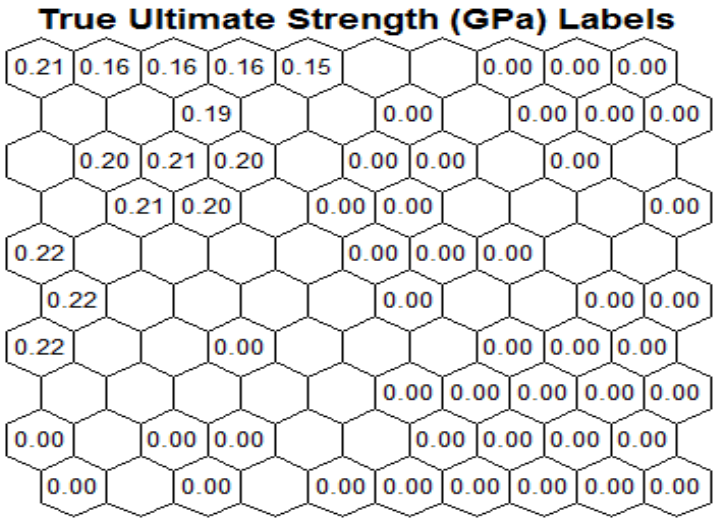


Figure 6.6 A 10×10 SOM based on the true ultimate strength response. The values are rounded for simplicity.

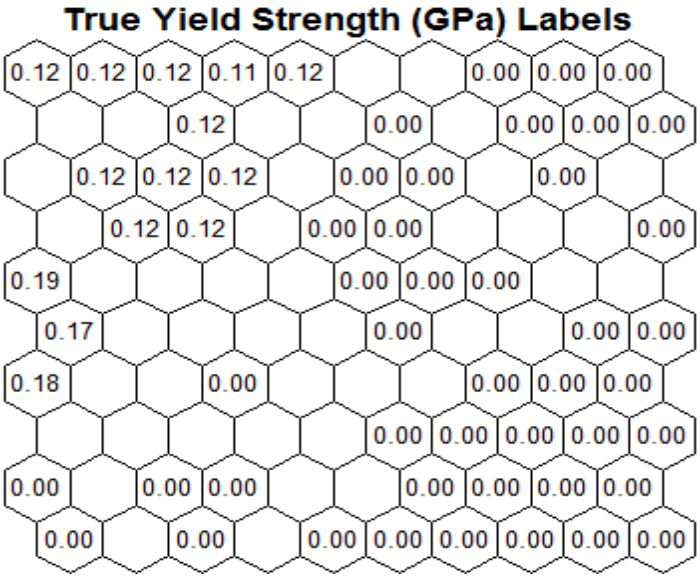


Figure 6.7 A 10×10 SOM based on the true yield strength response.

The values are rounded for simplicity.

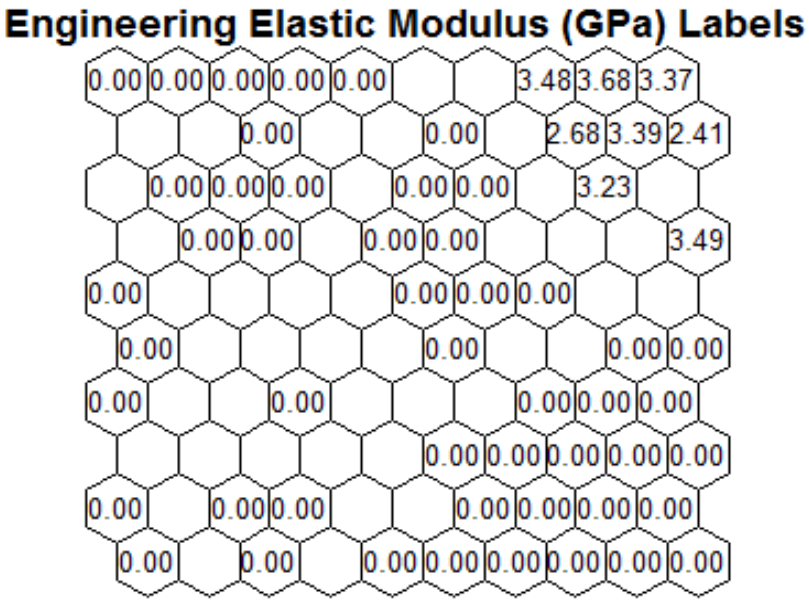


Figure 6.8 A 10×10 SOM based on the engineering elastic modulus response.

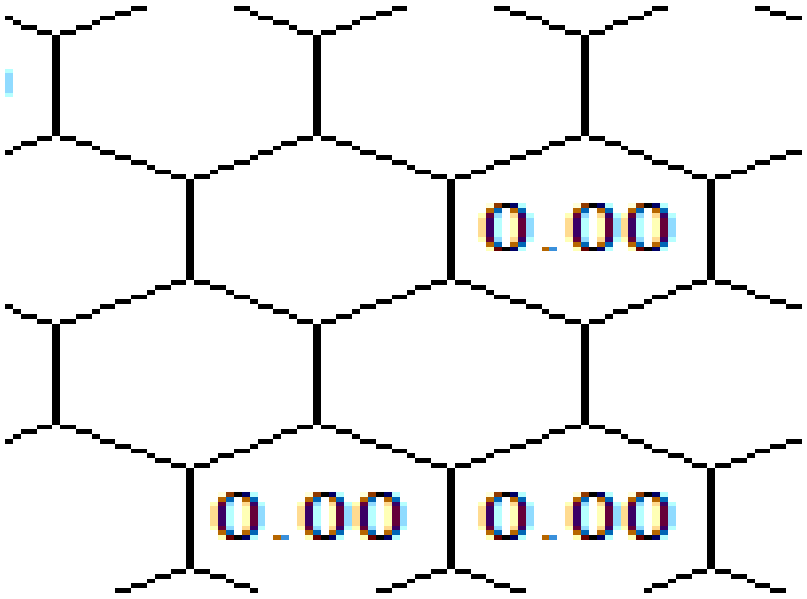


Figure 6.9 A 10×10 SOM based on the engineering ultimate strength response.

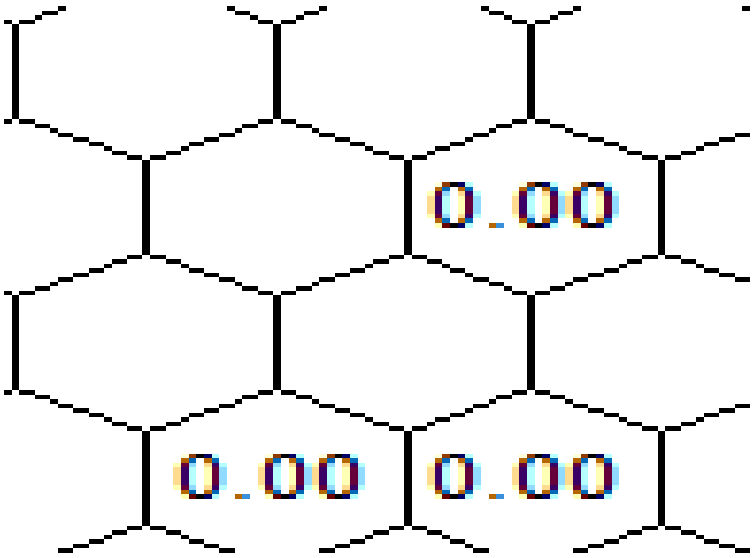


Figure 6.10 A 10×10 SOM based on the flexural modulus response.

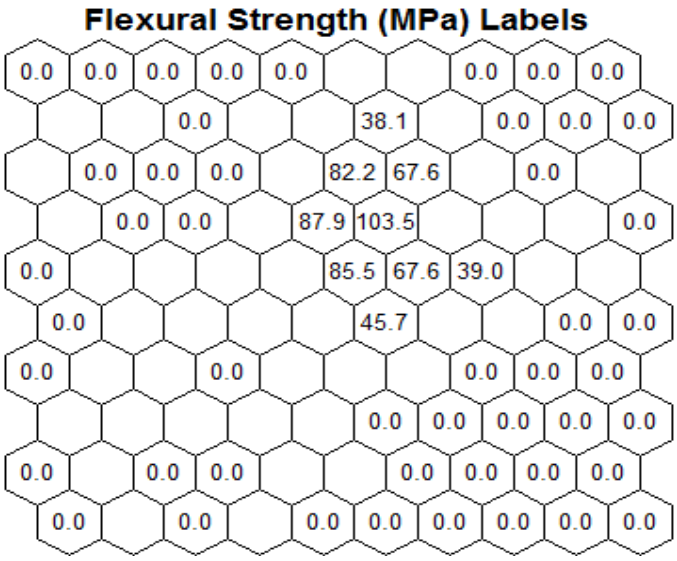


Figure 6.11 A 10×10 SOM based on the flexural strength response.

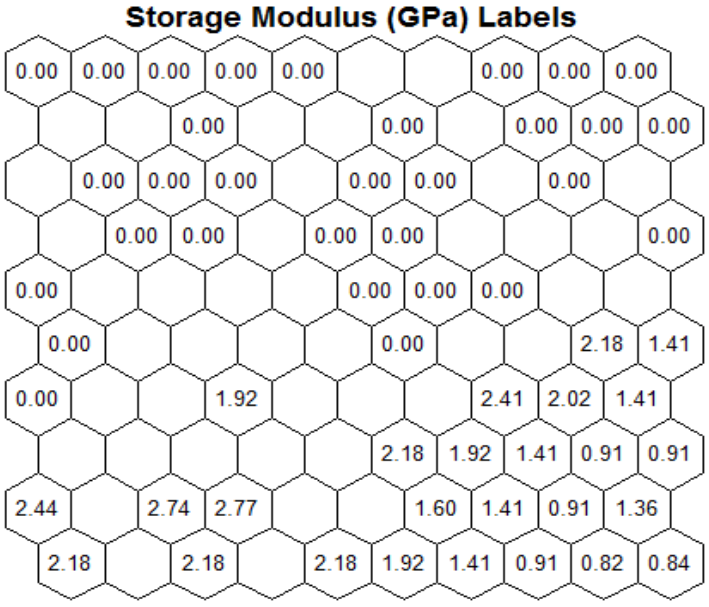


Figure 6.12 A 10×10 SOM based on the storage modulus response.

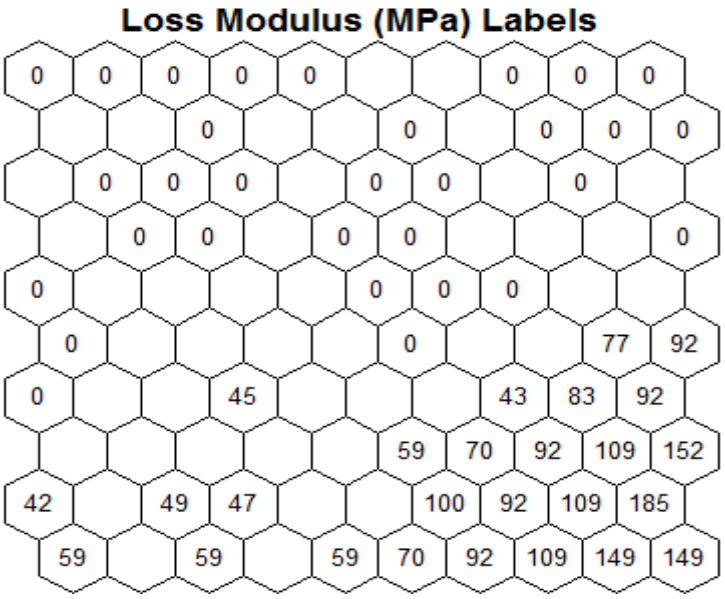


Figure 6.13 A 10×10 SOM based on the loss modulus response.
The values are rounded to the nearest integer for simplicity.

In addition to the sensitivity analysis inferred from SOMs, the different conditions needed to produce a particular optimal (highest) response can also be determined. In Figure 6.14, a 10×10 SOM is shown indicating the indices, which represent the numeric orders of the specimens mapped. Each index corresponds to one treatment combination out of 565 with specific values of curing environment, VGCNF mixing method, the presence or absence of a dispersing agent, fiber loading, VGCNF type, high shear mixing time, sonication time, temperature, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta. The indices in Figure 6.14 can be used to extract information linking the different dimensional combinations that produce the optimal response values.

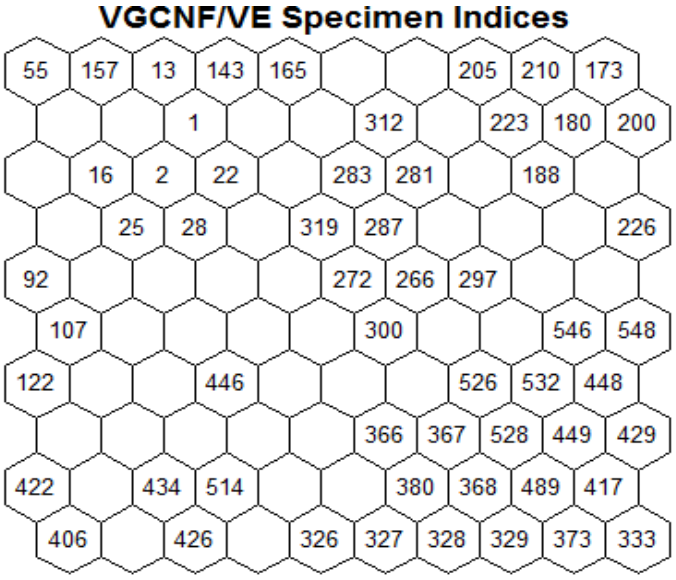


Figure 6.14 A 10×10 SOM illustrating the indices of the 565 nanocomposite specimens of the new framework [10, 13, 79].

From Figure 6.6, a group of three specimens have the highest true ultimate strength of about 0.22 GPa, located at the fifth, sixth, and seventh rows of the SOM. In Fig. 6.14, these values correspond to specimen indices 92, 107, and 122. Clearly, different input properties can be determined to produce the same 0.22 GPa response value. These properties are shown in Table 6.1. Nanocomposite designers can use such information in the selection of input factor levels.

Table 6.1 Different dimensional combinations required to produce an optimal true ultimate strength of about 0.22 GPa for the three specimens.

Optimal ultimate strength response value for the three specimens = 0.22 GPa	
Curing environment	Oxygen, Oxygen, Oxygen
Presence or absence of a DA	Yes, Yes, Yes
VGCNF mixing method	HS, US and HS, US and HS
Fiber loading (phr)	0.00, 0.50, 1.00
VGCNF type	Pristine, Pristine, Pristine
High shear mixing time (min)	0.00, 15.00, 15.00
Sonication time (min)	0.00, 60.00, 60.00
Temperature (^{o}C)	30.00, 30.00, 30.00

From Figure 6.7, one specimen has the highest true yield strength of about 0.19 GPa, located at the fifth row of the SOM. In Figure 6.14, this value corresponds to specimen index of 92 and different input properties can be determined to produce this response value. These properties are shown in Table 6.2.

Table 6.2 Different dimensional combinations required to produce an optimal true yield strength of about 0.19 GPa.

Optimal true yield strength response value = 0.19 GPa	
Curing environment	Oxygen
Presence or absence of a DA	Yes
VGCNF mixing method	HS
Fiber loading (phr)	0.00
VGCNF type	Pristine
High shear mixing time (min)	0.00
Sonication time (min)	0.00
Temperature (^{o}C)	30.00

From the sensitivity analysis that was conducted based on SOMs above, it was confirmed that temperature, sonication time, and high shear mixing time are the most dominant factors in the new nanocomposites framework. Therefore, one can focus on the quantities of these factors when determining the optimal conditions required to achieve

the optimal response value(s). From Table 6.1, high shear mixing time can be 0.00 or 15.00 minutes and sonication time can be 0.00 or 60 minutes and temperature must be at 30^{o}C. However, since the specimen of index 92 achieves both the optimal values of true ultimate strength and true yield strength, then very low values of high shear mixing and sonication times must be used at the testing temperature of 30^{o}C (Tables 6.1 and 6.2).

From Figure 6.8, one specimen has the highest engineering elastic modulus of about 3.68 GPa, located at the first row of the SOM. In Figure 6.14, this value corresponds to specimen index of 210 and different input properties can be determined to produce this response value. These properties are shown in Table 6.3.

Table 6.3 Different dimensional combinations required to produce an optimal engineering elastic modulus of about 3.68 GPa.

Optimal engineering elastic modulus response value = 3.68 GPa	
Curing environment	Nitrogen
Presence or absence of a DA	Yes
VGCNF mixing method	HS
Fiber loading (phr)	0.00
VGCNF type	Oxidized
High shear mixing time (min)	0.00
Sonication time (min)	0.00
Temperature (^{o}C)	30.00

From Figure 6.9, one specimen has the highest engineering ultimate strength of about 80.20 MPa, located at the first row of the SOM. In Figure 6.14, this value corresponds to specimen index of 173 and different input properties can be determined to produce this response value. These properties are shown in Table 6.4.

Table 6.4 Different dimensional combinations required to produce an optimal engineering ultimate strength of about 80.20 MPa.

Optimal engineering ultimate strength response value = 80.20 MPa	
Curing environment	Nitrogen
Presence or absence of a DA	No
VGCNF mixing method	US
Fiber loading (phr)	0.00
VGCNF type	Oxidized
High shear mixing time (min)	0.00
Sonication time (min)	0.00
Temperature (°C)	30.00

From Tables 6.3 and 6.4, in order to produce specimens with optimal values of engineering elastic modulus and engineering ultimate strength, the high shear mixing time and the sonication time must be very low and the testing temperature of 30°C must be used.

From Figure 6.10, one specimen has the highest flexural modulus of about 3.69 GPa, located at the fourth row of the SOM. In Figure 6.14, this value corresponds to specimen index of 319 and different input properties can be determined to produce this response value. These properties are shown in Table 6.5.

Table 6.5 Different dimensional combinations required to produce an optimal flexural modulus of about 3.69 GPa.

Optimal flexural modulus response value = 3.69 GPa	
Curing environment	Oxygen
Presence or absence of a DA	Yes
VGCNF mixing method	HS
Fiber loading (phr)	0.75
VGCNF type	Oxidized
High shear mixing time (min)	61.01
Sonication time (min)	20.47
Temperature (°C)	30.00

From Figure 6.11, one specimen has the highest flexural strength of about 103.5 MPa, located at the fourth row of the SOM. In Figure 6.14, this value corresponds to specimen index of 287 and different input properties can be determined to produce this response value. These properties are shown in Table 6.6.

Table 6.6 Different dimensional combinations required to produce an optimal flexural strength of about 103.5 MPa.

Optimal flexural strength response value = 103.5 MPa	
Curing environment	Oxygen
Presence or absence of a DA	Yes
VGCNF mixing method	HS
Fiber loading (phr)	0.25
VGCNF type	Pristine
High shear mixing time (min)	50.68
Sonication time (min)	8.64
Temperature (^{o}C)	30.00

From Tables 6.5 and 6.6, in order to achieve the optimal values of flexural modulus and flexural strength, generally low sonication time and relatively high values of high shear mixing time must be used. Testing temperature must be at 30^{o}C.

From Figure 6.12, two specimens have the highest storage modulus of about 2.76 GPa, located at the ninth row of the SOM. In Figure 6.14, this value corresponds to specimen indices of 434 and 514 and clearly different input properties can be determined to produce this response value. These properties are shown in Table 6.7.

Table 6.7 Different dimensional combinations required to produce an optimal storage modulus of about 2.76 GPa for the two specimens.

Optimal storage modulus response value for the two specimens = 2.76 GPa	
Curing environment	Oxygen, Oxygen
Presence or absence of a DA	Yes, Yes
VGCNF mixing method	HS, US and HS
Fiber loading (phr)	0.50, 0.50
VGCNF type	Pristine, Pristine
High shear mixing time (min)	28.87, 5.25
Sonication time (min)	36.80, 39.00
Temperature (^{o}C)	30.00, 30.00

From Figure 6.13, two specimens have the highest loss modulus of about 149 MPa, located at the tenth row of the SOM. In Figure 6.14, this value corresponds to specimen indices of 333 and 373 and clearly different input properties can be determined to produce this response value. These properties are shown in Table 6.8.

Table 6.8 Different dimensional combinations required to produce an optimal loss modulus of about 149 MPa for the two specimens.

Optimal loss modulus response value for the two specimens = 149 MPa	
Curing environment	Oxygen, Oxygen
Presence or absence of a DA	No, No
VGCNF mixing method	US, US
Fiber loading (phr)	0.25, 0.25
VGCNF type	Pristine, Oxidized
High shear mixing time (min)	66.15, 80.83
Sonication time (min)	25.91, 41.48
Temperature (^{o}C)	120.00, 120.00

From Tables 6.7 and 6.8, in order to produce specimens with optimal storage modulus values, the high shear mixing time must be low with moderately high sonication time and the testing temperature must be at 30^{o}C. On the hand, the optimal values of loss

modulus are obtained by using relatively high values of high shear mixing time and lower sonication time. The testing temperature in this case must be higher at 120^{o}C.

A PCA was run on the VGCNF/VE data in the newly designed nanocomposite framework. Figure 6.15 shows a graphical representation for the PCA of the data. PCA reduced the number of data dimensions from seventeen to two and each specimen was a given a specific 2-D representation (principal component 1 and 2 axes) so that specimens that have similar properties were mapped together in the 2-D space. Thus, there are no specific units associated with the abscissa and ordinate. This step is fundamental so that clustering algorithms (Section 3.4) can be applied to identify certain patterns in these nanocomposite data. Such patterns can be used to explain and discover certain physical/mechanical behavior associated with the data without running additional experiments.

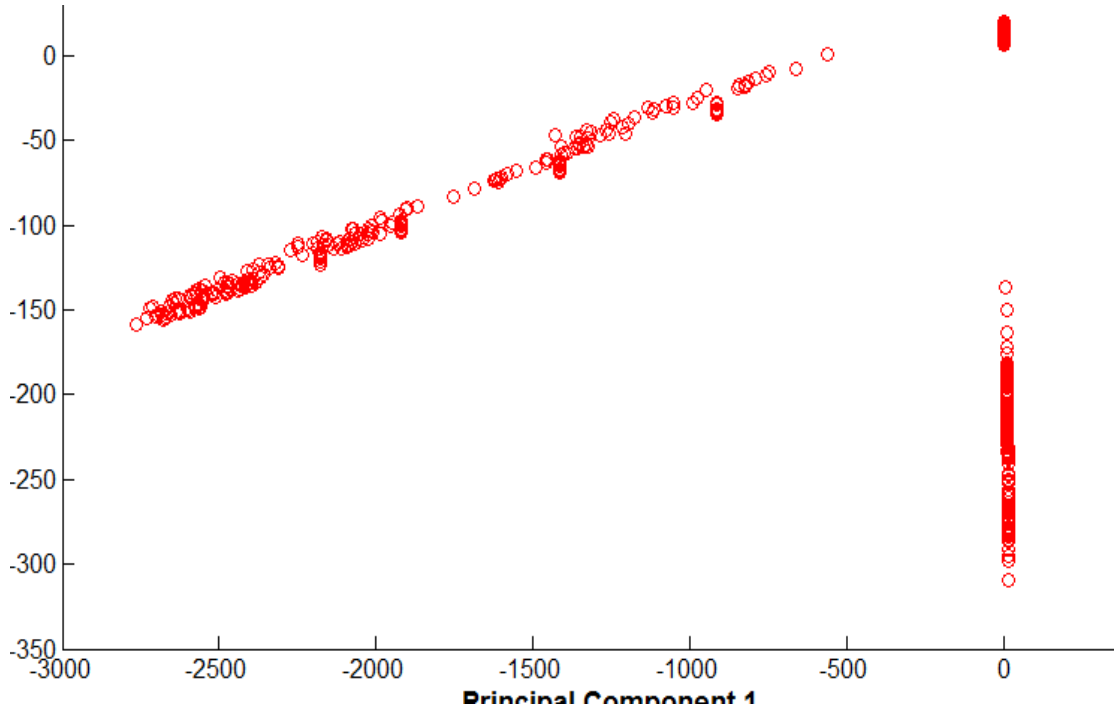
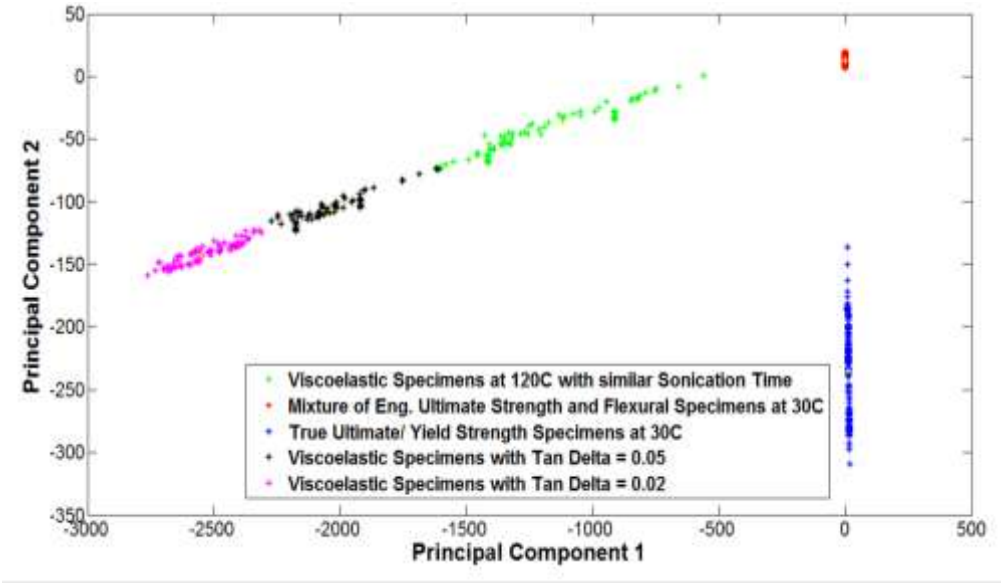


Figure 6.15 A 2-D graphical representation of the VGCNF/VE nanocomposite specimen data in the newly designed framework using the PCA technique.

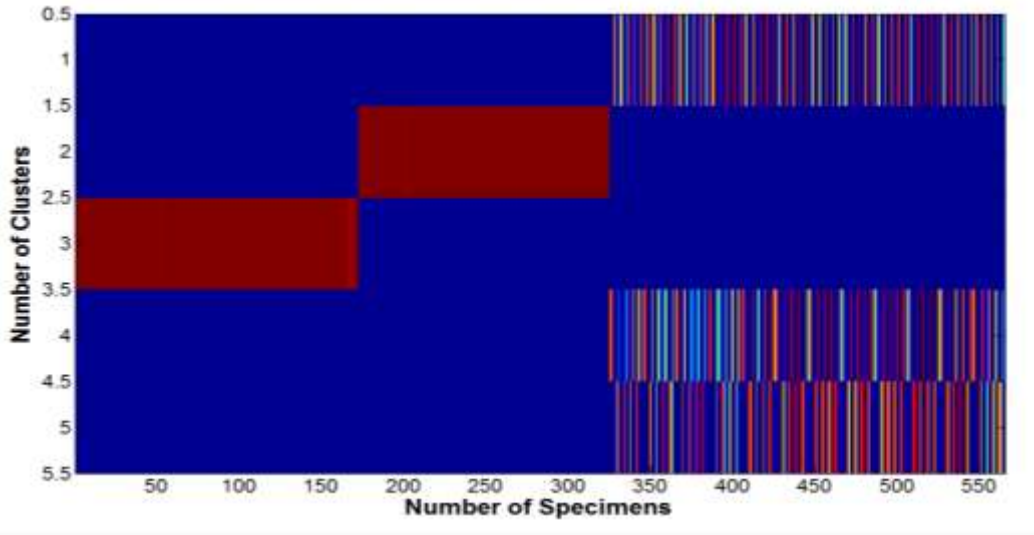
This technique maps the data from a 17-D space down to a 2-D space so that different clustering algorithms can be applied. The values associated with the principal dimensions 1 and 2 are random, but each specimen was given a 2-D coordinate so that specimens with similar properties would be mapped together in the 2-D space.

The FCM was applied to the new VGCNF/VE nanocomposite data using the GK distance measures. In Figure 6.16, the FCM results are illustrated, where five clusters are chosen to represent the data using the GK distance measure. The data points are divided into five different clusters, each shown with a different color. In Figure 6.16a, VGCNF/VE viscoelastic nanocomposite specimens were divided into three different clusters. The first one has viscoelastic specimens with the same tan delta value = 0.02. The second cluster has viscoelastic specimens with the same tan delta value = 0.05. The third cluster has viscoelastic specimens tested at 120^{o}C and have similar sonication times. This leads to the conclusion that VGCNF/VE viscoelastic specimens is important in the

newly designed nanocomposite framework as well as tan delta response is a dominant feature in this material system. In addition, the testing temperature of 120^{o}C and the sonication time are also important in the framework. Another observation that can be seen from Figure 6.16a is that specimens with engineering ultimate strength response and flexural data tested at 30^{o}C were clustered together in one cluster and specimens with true ultimate and yield strength responses tested at 30^{o}C were also placed in a separate cluster. This means that the testing temperature of 30^{o}C is a dominant feature in the framework as well as both flexural and impact strengths specimens are also vital in this material system. In addition, both engineering ultimate strength and flexural specimens have similar physical and mechanical behavior as they were placed in one cluster. In Figure 6.16b, a "scale data and display image (imagesc) object" plot is presented to indicate the number of clusters (each distinct set of bands in a row) and the bands associated with each cluster. The bands reflect the densities of data points within each cluster and correspond to the distances between the data points in Figure 6.16a. These findings prove that temperature is a dominant feature for the whole dataset.



a)



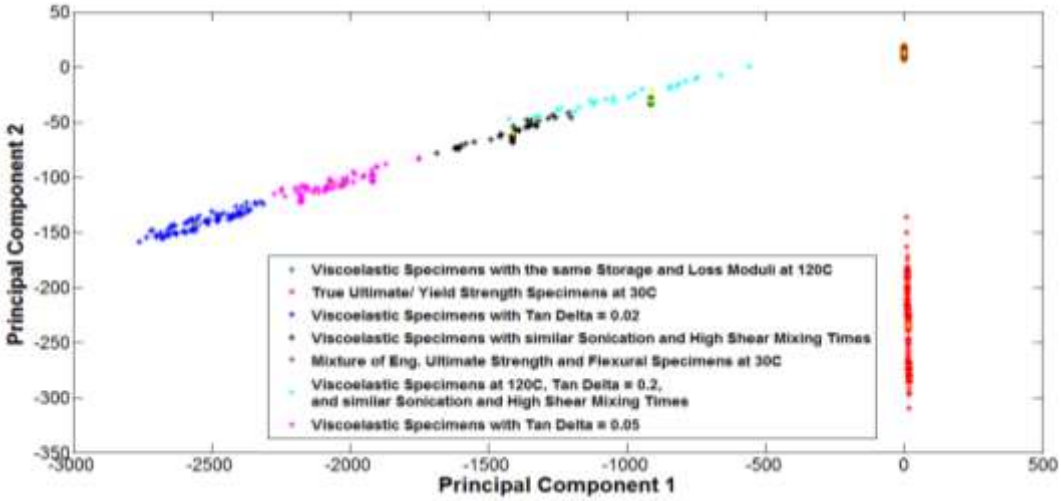
b)

Figure 6.16 Clustering results and imagesc plot after applying the FCM algorithm and the GK distance measure, when $C = 5$.

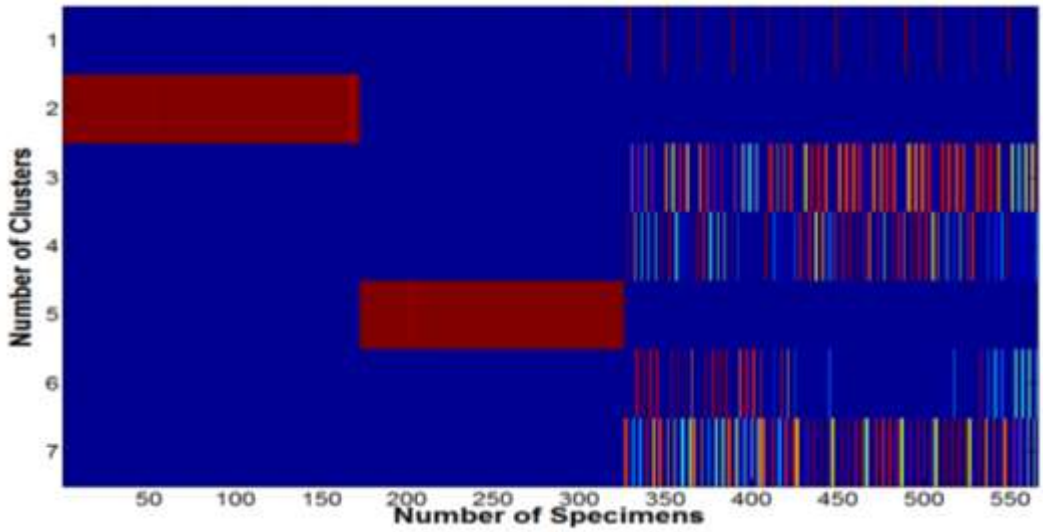
Clustering results are included in a). b) is the "scale data and display image (imagesc) object" plot where five bands representing five clusters can be identified.

In Figure 6.17, the FCM results are illustrated where seven clusters are chosen to represent the data using the GK distance measure. In Figure 6.17a, viscoelastic

VGCNF/VE specimens were divided into five different clusters. Particularly, one cluster with viscoelastic specimens that have the same tan delta response = 0.02, one cluster with viscoelastic specimens that have the same tan delta response = 0.05, one cluster with viscoelastic specimens with similar sonication and high shear mixing times, one cluster with viscoelastic specimens that have the same storage and loss moduli responses at the testing temperature of 120^{o}C, and one cluster with viscoelastic specimens tested at 120^{o}C with the same tan delta response value = 0.2 and have similar sonication and high shear mixing times. This leads to the conclusion that VGCNF/VE viscoelastic data is once again an essential component in the new framework. In addition, tan delta response is a dominant feature in this material system followed by the sonication time and high shear mixing time as some specimens were placed in some clusters based on their sonication and high shear mixing times. Moreover, the testing temperature of 120^{o}C is important in the framework as some viscoelastic specimens were placed in two clusters based on that temperature. In addition, similar to Figure 6.16a, specimens with engineering ultimate strength response and flexural data tested at 30^{o}C were clustered together in one cluster and specimens with true ultimate and yield strength responses tested at 30^{o}C were also placed in a separate cluster. This means that the testing temperature of 30^{o}C is a dominant feature in the framework as well as both flexural and impact strengths specimens are also important in this material system. In Figure 6.17b, an imagesc plot is presented, where seven clusters can be identified.



a)



b)

Figure 6.17 Clustering results and imagesc plot after applying the FCM algorithm and the GK distance measure, when $C = 7$.

Clustering results are included in a). b) is the "scale data and display image (imagesc) object" plot where seven bands representing seven clusters can be identified.

Using the GK distance measure, FCM works better for the 565 VGCNF/VE specimens when the selected number of clusters equals seven. For this case, specimens tested at different temperatures (particularly at 30°C and 120°C) and have the same tan

delta responses tend to be located in separate clusters that distinguish each of these temperatures and tan delta values. In addition, when the number of clusters equals to seven, more features that have pronounced effect in the new nanocomposite framework can be identified. For example, sonication time and high shear mixing time has come out to be important in the framework after applying FCM when seven clusters were selected. Also, viscoelastic specimens tested at 120^{o}C and have the same storage modulus and loss modulus responses have similar physical and mechanical behavior as they were placed in one separate cluster. These results confirmed some of the SOM findings above in that tan delta, temperature, sonication time, and high shear mixing time are the most dominant features in this material system and suggest that the FCM algorithm was able to identify VGCNF/VE specimens in the framework that have similar properties and placed them into different clusters based on tan delta, temperature, sonication time, high shear mixing time, and specimens type/ structure.

As mentioned in Chapter 4, clustering algorithms (e.g., FCM) can be used to better identify cogent patterns and trends in VGCNF/VE data. In addition, different VGCNF/VE specimens and their associated viscoelastic, flexural, and impact strengths properties can be identified and categorized within their respective clusters. Each cluster can be identified based on one or more of the input design factors of the VGCNF/VE material system in the newly designed nanocomposite framework.

CHAPTER VII

CONCLUSIONS AND FUTURE WORK

Signal processing and knowledge discovery techniques were applied to a vapor-grown carbon nanofiber (VGCNF)/vinyl ester (VE) nanocomposite dataset as a case study to validate different supervised and unsupervised learning approaches. This dataset had been generated by a full factorial experimental design with 240 different design points. Each treatment combination in the design consisted of eight feature dimensions corresponding to the design factors, i.e., VGCNF type, use of a dispersing agent, mixing method, VGCNF weight fraction, and testing temperature as the inputs and storage modulus, loss modulus, and tan delta as the output responses. Self-organizing maps (SOMs) were created with respect to temperature, tan delta, VGCNF weight fraction, storage modulus, and loss modulus. After analyzing the SOMs, temperature was identified as the dominant feature for the VGCNF/VE nanocomposites having the highest impact on the viscoelastic material responses. VGCNF weight fraction was also a dominant feature. In addition, it was inferred from the SOMs that some specimens tested at the same temperature tended to have several sub-clusters. Each sub-cluster had the same tan delta or VGCNF weight fraction values. Analyzing the SOMs with respect to storage and loss moduli demonstrated that VGCNF/VE specimens with different features could be designed to match a desired storage and/or loss modulus.

Another data analysis was performed using the principle component analysis (PCA) technique. Then, the fuzzy C-means (FCM) algorithm with the Gustafon-Kessel (GK) distance measure was applied to the resulting new dataset. The FCM clustered the specimens based on temperature as well as tan delta values. In addition, the FCM was able to recognize neat VE specimens tested at 30, 60, 90, and 120^{o}C and placed most of them in one cluster. In other words, when four clusters were selected and the GK distance measure was applied, neat VE specimens tested at 60-120^{o}C were placed in one cluster. In contrast, when five clusters were selected and the GK distance measure was applied, neat VE specimens tested at 90 and 120^{o}C were placed in one cluster. This reflects the fact that the viscoelastic properties of each neat VE specimen in both groups are similar. However, the FCM algorithm worked better when the number of clusters equals four, because more neat VE specimens tend to cluster together at the selected temperatures.

On the other hand, when VGCNF/VE specimens tested only at 60 and 90^{o}C used in the analysis, SOM was able to separate the specimens into two groups; one for 60^{o}C and the other for 90^{o}C specimens. Furthermore, SOM was able to identify subgroups within each group. That is, VGCNF/VE specimens with similar tan delta values tend to cluster together in the map to form subclusters within both 60 and 90^{o}C groups. In addition, the FCM was able to recognize neat VE specimens tested at 60, 90^{o}C and placed most of them in one cluster. In other words, when four clusters were selected and the GK distance measure was applied, neat VE specimens tested at 60, 90^{o}C were placed in one cluster. The other three clusters have specimens with the same tan delta values and/or the same testing temperature (60 or 90^{o}C). This is another proof that temperature and tan delta are dominant features in the VGCNF/VE dataset.

Next, Artificial Neural Network (ANN) technique was applied to VGCNF/VE nanocomposite dataset as another proof of concept for materials informatics. The ANN was trained using the resubstitution method and the 3-folds cross validation (CV) technique to provide a predictive model for these responses when the inputs are fed to the ANN. The ANN was able to predict/model these responses with minimal mean square error (MSE) using both techniques. However, the MSE error was relatively lower in case of resubstitution method. This is due to the fact that more samples were used for training and testing when the resubstitution method was implemented. In the 3-folds CV technique, the dataset was split into two subsets: one for training and one for testing (validation), so the number of samples used for training and testing was lower. This came at the expense of more converging time (23 epochs) needed for the ANN when the resubstitution method was implemented.

Furthermore, this work was extended to include other nanocomposite materials. The new dataset had been generated by a full factorial experimental design with 565 different design points representing three different nanocomposite structures, VGCNF/VE viscoelastic data, flexural data, and impact strengths data. Each treatment combination in the design consisted of seventeen feature dimensions corresponding to the design factors, i.e., curing environment, use or absence of dispersing agent, mixing method, VGCNF fiber loading, VGCNF type, high shear mixing time, sonication time and testing temperature were utilized as inputs and the true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus, loss modulus, and tan delta were selected as outputs. SOMs were created with respect to temperature, tan delta, high shear mixing time,

sonication time, VGCNF fiber loading, true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus and loss modulus. After analyzing the SOMs, temperature and tan delta were identified as the most dominant features for the newly designed VGCNF/VE nanocomposites framework having the highest impact on the material responses in the framework. Sonication time and high shear mixing time were also important. In addition, it was inferred from the SOMs that some specimens tested at the same temperature tended to have several sub-clusters. Each sub-cluster had similar tan delta values. The cluster with the highest number of specimens in the "temperature labels" SOM is the 30^{o}C cluster. This means that 30^{o}C is the most important temperature as it drives the behavior of all specimens in the newly designed framework. Analyzing the SOMs with respect to true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus and loss modulus demonstrated that VGCNF/VE specimens with different features could be designed to match an optimal value of VGCNF/VE impact strength response and/or VGCNF/VE flexural response and/or VGCNF/VE viscoelastic response.

Finally, another data analysis was performed using the principle component analysis PCA technique. Then, FCM algorithm with GK distance measure was applied to the resulting new dataset. The FCM clustered the specimens based on temperature, tan delta values as well as sonication time and high shear mixing time. However, the testing temperature of 30^{o}C and 120^{o}C were the most important temperatures as specimens were clustered based on these two particular temperatures. In addition, the FCM was able to recognize the viscoelastic specimens tested at 120^{o}C and have the same storage and loss

modulus values and placed them in one cluster. This reflects the fact that the mechanical and physical properties of these specimens are similar. In addition, when seven clusters were selected and the GK distance measure was applied, there was one cluster that had only VGCNF/VE impact strengths specimens, one cluster that had a mixture of engineering ultimate strength and flexural specimens, and five clusters that had VGCNF/VE viscoelastic specimens with different properties. This means that all nanocomposites structures in the framework are important and the VGCNF/VE viscoelastic specimens are the most important structure. Moreover, the FCM algorithm worked better when the number of clusters equals seven, because high shear mixing time came out to be an important feature in the new framework. In addition, when the number of clusters equal seven, the viscoelastic specimens that have the same storage and loss moduli tend to be placed in one cluster.

In summary, based on the above, the main contribution of this study is to design a data analytic materials informatics methodology for nanocomposites that consists of the following signal processing approaches:

- Developing a sensitivity analysis structure using SOMs in order to discover the most and least dominant features of the VGCNF/VE system, whether they are input design factors or output responses.
- Developing a tool for identifying VGCNF/VE specimen designs leading to the same storage and loss moduli. This will facilitate tailoring of nanocomposite viscoelastic properties and, in turn, minimize fabrication costs by the domain experts.

- Developing a methodology to better identify cogent patterns and trends in VGCNF/VE data. Each cluster can be identified based on one or more of the input design factors of the VGCNF/VE system.
- Developing a procedure where the output responses of VGCNF/VE system can be modeled based on the input design factors using ANN algorithms and techniques like the back-propagation (BP) algorithm and the *n*-folds cross validation technique.
- Developing a tool for identifying VGCNF/VE specimen designs leading to the optimal (highest) responses of true ultimate strength, true yield strength, engineering elastic modulus, engineering ultimate strength, flexural modulus, flexural strength, storage modulus and loss modulus. This will facilitate tailoring of nanocomposite viscoelastic, impact strengths, and flexural properties and, in turn, minimize fabrication costs and increase the production efficiency by the domain experts.

The signal processing, knowledge discovery, and other supervised and unsupervised learning approaches applied here demonstrate the dominant features in the nanocomposite data without the need to conduct additional expensive and time-consuming experiments. This not only validates these approaches but also highlights the feasibility of data mining and knowledge discovery techniques in materials science and engineering and the rising field of Materials Informatics.

The effectiveness of data mining and knowledge discovery techniques in engineering fields makes this a rich topic with many possible avenues of investigation. In particular, the future agenda is not limited to validating and discovering certain trends,

properties, and behavior related to particular system/application (although this certainly remains a challenge), but encompasses the proper collaboration with the domain experts in order to work on the datasets that are difficult to be modeled using other physical, mathematical or statistical models.

The applications of this research will be numerous and diverse, for the simple reason that signal processing and intelligent systems have immediate relevance and impact to the design and cost effectiveness of manufacturing processes. Also it is important to understand how this research can expand and grow as well as how it can develop into new topics. The remainder of this chapter briefly describes some specific potential research projects.

The first topic supports some continuation of this research, while the second topic expands on this research. The third and fourth topics will allow expanding the application of this research to a completely new area of study. Briefly, the four research areas are provided below:

- **Combining (fusing) multiple datasets:** This can be accomplished by establishing a baseline by which different datasets from multiple domains, related to different applications in the same domain, can be combined (fused) in order to discover new trends and characteristics of the new applications as well as to validate the facts mentioned in theory without the need to conduct expensive and time-consuming experiments.
- **Designing 3G AHSS as a new original application:** Currently, there are not enough samples (data points) in the 1st and 2nd generation dataset. In addition, as received from the domain experts in this area, there is no

correlation at all between the features currently used (i.e. between the chemical compositions and the ultimate tensile strength, UTS) because UTS relates more to microstructure *not* to chemical compositions except for elastic structures. Thus, In order to design a framework for 3G AHSS, the features have to be modified and more samples have to be created (collected).

- **Designing an ontology-based framework:** The materials science ontology, represented in the Web Ontology Language (OWL), enables (1) the mapping between and the integration of different materials science databases, (2) the modeling of experimental provenance information acquired in the physical and digital domains and, (3) the inferencing and extraction of new knowledge within the materials science domain. Consequently, this framework will enable the scientists to search, retrieve, correlate and integrate diverse, but related, domain-specific data and information across different databases.
- **Applying the proposed methodology onto other materials systems**: In addition to the three nanocomposites structures this methodology utilized, it can be further validated and extended using not only other nanocomposites and polymers structures, but also other materials systems like metals, steels, ceramics,…etc.

While the research areas discussed above will address academia issues, it should be firmly believed that industrial problems should also be answered by this research. As such, future researchers will have to aggressively seek to form partnerships with industry

to answer their pragmatic needs, which could also provide additional funding support for their research topics.

REFERENCES

[1] X. Yao, "Research Issues in Spatio-temporal Data Mining," A white paper submitted to the University Consortium for Geographic Information Science (UCGIS) workshop on Geospatial Visualization and Knowledge Discovery, Lansdowne, Virginia, Nov. 18-20, 2003.

[2] Miller, H.J. and Han, J. (2001), Geographic data mining and knowledge discovery: an overview. In Miller, H.J. and Han, J. (eds) *Geographic data mining and knowledge discovery*. London, New York : Taylor & Francis, 3-32.

[3] Fayyad, U.M., Piatetsky-Shapiro, G. Smyth, P. (1996), From data mining to knowledge discovery: An Overview. In Fayyad, U.M., Piatetsky-Shapiro, G. Smyth, P. Ulthurusamy, R. (eds) *Advances in Knowledge Discovery and Data Mining*. Cambridge, MA:MIT Press, 1-34.

[4] Durbha, S.S. and King, R.L., Semantics-Enabled Framework for Knowledge Discovery from Earth Observation Data Archives, *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 43, Issue 11, pp. 2563 – 2572, November 2005.

[5] Durbha, S.S., R.L. King, and N.H. Younan, An Information Semantics Approach for Knowledge Management and Interoperability for the Global Earth Observation System of Systems, *IEEE Systems Journal* (Special Issue on GEOSS), Vol. 2, No. 3, pp. 358-365, September 2008.

[6] J. H. Koo, Polymer nanocomposites: Processing, characterization, and applications, first edition, McGraw-Hill, New York, 2006.

[7] J. Garces, D. J. Moll, J. Bicerano, R. Fibiger, D. G. McLeod, Polymeric nanocomposites for automotive applications, J. Adv. Mater.. 12 (2000) 1835-1839.

[8] F. Hussain, M. Hojjati, M. Okamoto, R. E. Gorga, Review article: polymer-matrix nanocomposites, processing, manufacturing, and application: An overview, J. Compos. Mater.. 40 (2006) 1511-1575.

[9] E. T. Thostenson, C. Li, T. W. Chou, Nanocomposites in context, J. Compos. Sci. Technol.. 65 (2005) 491-516.

[10] S. Nouranian, Vapor-grown carbon nanofiber/vinyl ester nanocomposites: Designed experimental study of mechanical properties and molecular dynamics simulations, PhD Dissertation, Mississippi State University, Mississippi State, MS, USA, 2011.

[11] S. Nouranian, H. Toghiani, T. E. Lacy, C. U. Pittman, J. Dubien, Dynamic mechanical analysis and optimization of vapor-grown carbon nanofiber/vinyl ester nanocomposites using design of experiments, J. Compos. Mater.. 45 (2011) 1647-1657.

[12] S. Nouranian, T. E. Lacy, H. Toghiani, C. U. Pittman Jr, J. L. Dubien, Effects of Formulation, Processing, and Temperature on the Viscoelastic Properties of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites, J. Applied Polymer Sci., 2012 (Submitted).

[13] G. W. Torres, S. Nouranian, T. E. Lacy, H. Toghiani, C. U. Pittman Jr, J. Dubien, Statistical Characterization of the Impact Strengths of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites Using a Central Composite Design, J. Applied Polymer Sci.. First published online (2012) DOI: 10.1002/app.38190.

[14] G. G. Tibbetts, M. L. Lake, K. L. Strong, B. P. Rice, A review of the fabrication and properties of vapor-grown carbon nanofiber/polymer composites, J. Compos. Sci. Technol.. 67 (2007) 1709-1718.

[15] A. Plaseied, A. Fatemi, M. R. Coleman, Influence of Carbon Nanofiber Content and Surface Treatment on Mechanical Properties of Vinyl Ester, J. Polym. Polym. Compos.. 16 (2008) 405-413.

[16] A. Plaseied, A. Fatemi, Tensile creep and deformation modeling of vinyl ester polymer and its nanocomposite, J. Reinforced Plastics Compos.. 28 (2009) 1775-1788.

[17] Qunyi, Xiaodong, Xiangguo, Weidong, Materials informatics and study on its further development. *Chinese Science Bulletin* 2006 Vol. 51 No. 4 498-504.

[18] Toyohiro Chikyow.: (2006) Trends In Materials Informatics In Research On Inorganic Materials. *Quarterly Review* , Vol.20, pp. 59-71.

[19] Rajan, K., Materials Informatics: An Introduction. CoSMIC, Iowa State University. Materials Technology@TMS, 2007.

[20] Hu, C, Ouyang, C, Wu, J, Zhang, X, Zhao, C, Non-Structured Materials Science Data Sharing Based on Semantic Annotation. *Data Science Journal*, Volume 8, 20 May 2009.

[21] Paolini, C, Bhattacharjee, S, A Web Service Infrastructure for Thermochemical Data. *J. Chem. Inf. Model.* 2008, 48, 1511-1523

[22] Chemical Abstracts Service (CAS) Registry. http://www.cas.org/ (accessed April, 7, 2011)

[23] Sabin, T J, Bailer-Jones, C A L, Withers, P J, Accelerated learning using Gaussian process models to predict static recrystallization in an Al-Mg alloy. *Modelling Simul. Mater. Sci. Eng*. 8 (2000) 687-706

[24] Roberts, K, Muchlich, F, Schenkel, R, Weikum, G, An Information System for Material Microstructures. *International Conference on Scientific and Statistical Database Management (SSDBM'04)*. 1099-3371/04, 2004 IEEE

[25] Pouchard, L, Rana, O, Walker, D. An Ontology for User Support on the Materials Microcharacterization Collaboratory. *Proceedings of 5th International Conference on Autonomous Agents*. Montreal, Canada. 2001.

[26] Belov, G, Iorish, V. T. The Data Exchange Formats on Thermodynamic Properties of Individual Substances. IHED, IVTAN Association of RAS, Thermocenter, Izhorskaya 13/19, Moscow, Russia, gbelov@imail.ru.

[27] Guessasma, S, Coddet, C. Microstructure of APS Alumina-Titania Coatings Analyzed Using Artificial Neural Network. *Acta Materialia* 52 (2004) 5157-5164.

[28] Swaddiwudhipong, S, Tho, K K, Liu, Z S, Hua, J, Ooi, N S B. Material Characterization via Least Squares Support Vector Machines. *Modelling Simul. Mater. Sci. Eng*. 13 (2005) 993-1004.

[29] Huber, N, Tsagrakis, I, Tsakmakis, Ch. Determination of constitutive properties of think metallic films on substrates by spherical indentation using neural networks. *Int. J. Solids Struct.* 37 (2000) 6499-516.

[30] Huber, N, Nix, W D, Gao, H. Identification of elastic-plastic material parameters from pyramidal indentation of thin films. *Proc. R. Soc. Lond*. (2000) A 458 1593-620.

[31] Swaddiwudhipong, S, Tho, K K, Liu, Z S, Zeng K. Material characterization based on dual indenters. *Int. J. Solids Struct.* (2005) 42 69-83.

[32] Pelckmans, K, Suykens, J A K, Van Gestel, T, De Brabanter, J, Lukas, L, Hamers, B, De Moor, B, Vandewalle, J. LS-SVMlab: a Matlab/C toolbox for least squares support vector machines. *Internal Report* (2002) 02-44, ESAT-SISTA (Belgium: K.U. Leuven)

[33] Mukherjee, M, Singh, S, Mohanty, O. Neural network analysis of strain induced transformation behaviour of retained austenite in TRIP-aided steels. *Materials Science and Engineering* A 434 (2006) 237-245

[34] Bhadeshia, H.K.D.H, MacKay, D. J.C, Svensson, L-E. The Impact Toughness of C-Mn Steel Arc-Welds- A Bayesian Neural Network Analysis. *Materials Science and Technology* 11 (1995) 1046-1051.

[35] Yescas, M.A., Bhadeshia, H.K.D.H. Model for the maximum fraction of retained austenite in austempered ductile cast iron. *Materials Science and Engineering* A333 (2002) 60-66.

[36] Yoshitake, S., Narayan, V., Harada, H.,Bhadeshia, H.K.D.H., Mackay, D.J.C. Estimation of the γ and γ' Lattice Parameters in Nickel-base Superalloys Using Neural Network Analysis. *ISIJ International*, Vol. 38 (1998), No. 5, pp. 495-502.

[37] Tancret, F., Bhadeshia, H.K.D.H, Mackay, D.J.C. Design of a creep resistant nickel base superalloy for power plant applications, Part 1- Mechanical properties modeling. *Materials Science and Technology*. March 2003, Vol.19.

[38] Wei, Y. Bhadeshia, H.K.D.H, Sourmail, T. Mechanical Property Prediction of Commercially Pure Titanium Welds with Artificial Neural Network. *Materials Science and Technology*, Vol. 21, No. 3, 2005.

[39] Singh, S.B., Bhadeshia, H.K.D.H, MacKay, D.J.C, Carey, H., Martin, I. Neural network analysis of steel plate processing. *Ironmaking and Steelmaking* 1998, Vol. 25, No.5.

[40] Badmos, A.Y., Bhadeshia, H.K.D.H, MacKay, D.J.C. Tensile properties of mechanically alloyed oxide dispersion strengthened iron alloys- Part 1- Neural network models. *Materials Science and Technology*, August 1998, Vol. 14.

[41] Narayan, V., Abad, R., Lopez, B., Bhadeshia, H.K.D.H, MacKay, D.J.C. Estimation of Hot Torsion Stress Strain Curves in Iron Alloys Using a Neural Network Analysis. *ISIJ International*, Vol.39, No.10, pp. 999-1005.

[42] Ryu, Joo, Bhadeshia, H.K.D.H. Contribution of Microalloying to the Strength of hot-rolled Steels. *Materials and Manufacturing Processes* 24 (2009) 1-7.

[43] Gavard, L., Bhadeshia, H.K.D.H, MacKay, D.J.C., Suzuki, S. Bayesian neural network model for austenite formation in steels. *Materials Science and Technology*, June 1996, Vol.12.

[44] Kemp, R., Cottrell, G.A., Bhadeshia, H.K.D.H. Designing Optimized Experiments for the International Fusion Materials Irradiation Facility. *Journal of Nuclear Materials*, Volumes 367-370, 2007, 1586-1589.

[45] Fujii, Hidetoshi, MacKay, D.J.C., Bhadeshia, H.K.D.H. Bayesian Neural Network Analysis of Fatigue Crack Growth Rate in Nickel Base Superalloys. *ISIJ International*, Vol. 36 (1996), No.11, pp. 1373-1382.

[46] Forsik, S., Bhadeshia, H.K.D.H. Elongation of Irradiated Steels. *Materials and Manufacturing Processes*, Vol. 24 (2009), 130-137.

[47] Cottrell. G. A., Kemp, R., Bhadeshia, H.K.D.H., Odette, G.R., Yamamoto, T., Neural Network Analysis of Charpy Transition Temperature of Irradiated Low-activation Martensitic Steels. *Journal of Nuclear Materials*, Volumes 367-370, 2007, Pages 603-609.

[48] Pak, J, Jang, J, Bhadeshia, H.K.D.H., Karlsson, L. Optimization of Neural Network for Charpy Toughness of Steel Welds. *Materials and Manufacturing Processes* 24 (2009) 16-21.

[49] Sourmail, T., Bhadeshia, H.K.D.H., MacKay, D.J.C. Neural network model of creep strength of austenitic stainless steels. *Materials Science and Technology*, June 2002, Vol. 18.

[50] Chatterjee, S., Murugananth, Bhadeshia, K.K.D.H. δ TRIP steel. *Materials Science and Technology*. 2007, Vol. 23, No. 7

[51] Tancret, F., Bhadeshia, H.K.D.H., MacKay, D.J.C., Design of new creep-resistant nickel-base superalloys for power plant applications. *Key Engineering Materials*, Volumes 171-174, 2000, pp. 529-536.

[52] Das, S., Singh, S.B., Mohanty, O.N., Bhadeshia, H.K.D.H. Understanding the complexities of bake hardening. *Materials Science and Technology*, 2008, Vol. 24, No. 1.

[53] Brun, F., Yoshida, T., Robson, J.D., Narayan, V., Bhadeshia, H.K.D.H., MacKay, D.J.C. Theoretical design of ferritic creep resistant steels using neural network, kinetic, and thermodynamic models. *Materials Science and Technology*. May 1999, Vol. 15.

[54] Metzbower, E. A., DeLoach, J.J., Lalam, S., Bhadeshia,H.K.D.H. Secondary effects in neural network analysis of the mechanical properties of welding alloys for HSLA shipbuilding steels. *Mathematical Modelling of Weld Phenomena – VI*. Published by the Institute of Materials, eds Cerjak, H and Bhadeshia, H. K. D. H, 2002, 231-242

[55] Keehan, E., Andrén, H.O., Karlsson, L., Murugananth, M., Bhadeshia, H.K.D.H. Microstructural and Mechanical effects of nickel and manganese on high strength steel weld metals. *Trends in Welding Research,* eds David, S.A., Vitek, J., Debroy, T., Lippold, J., Smartt, H. ASM International, USA, 2002, 719-723.

[56] Dimitriu, R.C., Bhadeshia, H.K.D.H. Hot strength of creep resistant ferritic steels and relationship to creep rupture data. *Materials Science and Technology*. 2007, Vol 23, No. 9.

[57] Dimitriu, R.C., Bhadeshia, H.K.D.H., Fillon, C., Poloni, C., Hot-Strength of Ferritic Creep-Resistant Steels Comparison of Neural Network and Genetic Programming. *Materials and Manufacturing Processes*, Vol. 24 (2009) 10-15.

[58] A. Javadi, M. Rezania, Intelligent finite element method: An evolutionary approach to constitutive modeling, Advanced Engineering Informatics, 23 (2009) 442-451.

[59] I.K. Brilakis, L. Soibelman, Y. Shinagawa, Construction site image retrieval based on material cluster recognition, Advanced Engineering Informatics, 20 (2006) 443-452.

[60] A. Ullah, K.H. Harib, An intelligent method for selecting optimal materials and its application, Advanced Engineering Informatics, 22 (2008) 473-483.

[61] R. L. King, A. Rosenberger, L. Kanda, Artificial neural networks and three dimensional digital morphology: a pilot study, J. Folia Primatologica. 76 (2005) 303-324.

[62] T. Kohonen, Self-organization and associative memory, third ed. ,Springer-Verlag, New York, 1989.

[63] I. T. Jolliffe, Principal Component Analysis, second ed., Springer, New York, 2002.

[64] S. Miyamoto, H. Ichihashi, K. Honda, Algorithms for Fuzzy Clustering: Methods in c-Means Clustering with Applications, Springer, 2008.

[65] J.C. Bezdek, R. Ehrlich, FCM: The fuzzy c-means clustering algorithm, Computers & Geosciences, 10 (1984) 191-203.

[66] S. Nouranian, T.E. Lacy, H. Toghiani, C.U. Pittman Jr, J.L. Dubien, Response Surface Predictions of the Viscoelastic Properties of Vapor-Grown Carbon Nanofiber/Vinyl Ester Nanocomposites, J. Appl. Polym. Sci., (2013) DOI: 10.1002/app.39041.

[67] D.C. Montgomery, Design and analysis of experiments, 7th ed., John Wiley & Sons, Hoboken, NJ, 2009.

[68] J.W. Sammon Jr, A nonlinear mapping for data structure analysis, Computers, IEEE Transactions on, 100 (1969) 401-409.

[69] C. Ding and X. He, K-means Clustering via Principal Component Analysis, *Proceedings of the twenty-first international conference on Machine learning,* ACM, pp. 29-37, July 2004.

[70] S. Theodoridis and K. Koutroumbas, Pattern Recognition, fourth ed., Academic Press, Massachusetts, 2008.

[71] MATLAB help, version 7.14.0.739 (R2012a), February 09, 2012.

[72] J. Twomey, A. Smith, Bias and Variance of Validation Methods for Function Approximation Neural Networks Under Conditions of Sparse Data, IEEE Transactions on Systems, Man., and Cybernetics, 28 (1998) 417-430.

[73] R. S. Yassar, O. AbuOmar, E. Hansen, M. F. Horstemeyer, On dislocation-based artificial neural network modeling of flow stress, J. Mater. Design. 31 (2010) 3683-3689.

[74] S. Hayken, Neural Networks: A Comprehensive Foundation, 2nd ed., Englewood Cliffs (NJ): Prentice-Hall; 1999.

[75] S. Haykin, Neural Networks and Learning Machines, 3rd ed., Prentice-Hall, 2009.

[76] Matlock, D.K., Speer, J.G.: Microstructure and Texture in Steels and Other Materials, *A. Haldar, S. Suwas and D. Bhattachrajee Eds., Springer, London* (2009) pp. 185 – 204.

[77] J. D. Ferry, Viscoelastic Properties of Polymers, third ed., Wiley, New York, 1980.

[78] L. E. Nielsen and L. E. N. R. F. Landel, Mechanical Properties of Polymers Composites, second ed., Marcel Dekker, New York, 1994.

[79] J. Lee, S. Nouranian, G.W. Torres, T. E. Lacy, H. Toghiani, C. U. Pittman, J. L. DuBien, , Characterization, prediction, and optimization of flexural properties of vapor-grown carbon nanofiber/vinyl ester nanocomposites by response surface modeling. J. Appl. Polym. Sci., 130 (2013) 2087–2099. doi: 10.1002/app.39380

[80] MATLAB Mathematics and Interpolation, Release 2012a, The MathWorks, Inc., Natick, Massachusetts, United States.

[81] O. Abuomar, S. Nouranian, R. King, J.L. Bouvard, H. Toghiani, T. E. Lacy, C. U. Pittman Jr, Data mining and knowledge discovery in materials science and engineering: A polymer nanocomposites case study, J. Advanced Engineering Informatics, 27 (2013) 615–624