G. YAZICI

ATILIM UNIVERSITY 2022

PERFORMANCE EVALUATION OF PREPROCESSING TO PCA COMBINED MACHINE LEARNING TECHNIQUES ON PHARMACEUTICAL AND MINERAL SAMPLES BY LASER-INDUCED BREAKDOWN SPECTROSCOPY

ATILIM UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

GÖKTUĞ YAZICI

MASTER OF SCIENCE THESIS

THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

JANUARY 2023

PERFORMANCE EVALUATION OF PREPROCESSING TO PCA COMBINED MACHINE LEARNING TECHNIQUES ON PHARMACEUTICAL AND MINERAL SAMPLES BY LASER-INDUCED BREAKDOWN SPECTROSCOPY

ATILIM UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

GÖKTUĞ YAZICI

MASTER OF SCIENCE THESIS

THE DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

JANUARY 2023

Approval of the Graduate School of Natural and Applied Sciences, Atilim University.

_____

Prof. Dr. Ender KESKİNKILIÇ
Director of Graduate School

I certify that this thesis satisfies all the requirements as a thesis for the degree of **Master of Science in Electrical and Electronics Engineering, Atılım University**.

_____

Prof. Dr. Reşat Özgür DORUK
Head of Department

This is to certify that we have read the thesis PERFORMANCE EVALUATION OF PREPROCESSING TO PCA COMBINED MACHINE LEARNING TECHNIQUES ON PHARMACEUTICAL AND MINERAL SAMPLES BY LASER-INDUCED BREAKDOWN SPECTROSCOPY submitted by GÖKTUĞ YAZICI and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____ _____

Prof. Dr. Reşat Özgür DORUK
Supervisor

Examining Committee Members:

Prof. Dr. Halil Berberoğlu
Department of Physics
Ankara Hacı Bayram Veli University

_____

Prof. Dr. Reşat Özgür DORUK
Electrical and Electronics Eng.
Department, Atılım University

_____

Prof. Dr. Mehmet IŞIK
Physics Group,
Atılım University

_____

Date: 11/01/2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Göktuğ YAZICI

Signature :

ABSTRACT

PERFORMANCE EVALUATION OF PREPROCESSING TO PCA COMBINED MACHINE LEARNING TECHNIQUES ON PHARMACEUTICAL AND MINERAL SAMPLES BY LASER-INDUCED BREAKDOWN SPECTROSCOPY

YAZICI, Göktuğ

MSc., Department of Electrical and Electronics Engineering

Supervisor : Prof. Dr. Reşat Özgür DORUK

January 2023, 82 pages

For the purpose of identifying and analyzing materials, laser-induced breakdown spectroscopy (LIBS) is a quick optical nuclear discharge spectroscopy. It has the advantages of in-situ analysis, removal of rigorous sample processing, and micro-destructive properties for the substance being evaluated. LIBS uses brief bursts of laser beams to stimulate the material to a certain threshold, resulting in plasma formation. The plasma properties, which include wavelength value and intensity amplitude, are affected by the material and the surroundings of the experiment. The spectrum profiles of medication and mineral samples were obtained using LIBS in this study. The collection of pharmaceutical samples comprises two distinct concentrations of both paracetamol-based drugs, Aferin and Parafon. Aluminum (Al), Bizmut (Bi), Copper (Cu), Iron (Fe), Manganese (Mn), Nickel-Aluminum (NiAl), Tin (Sn), and Zinc (Zn) are among the mineral samples in the dataset. The samples' spectrum data were preprocessed by replacing missing values with shape-preserving piecewise cubic spline interpolation, filling outliers based on quartiles, smoothing spectra to remove noise, and normalizing both the wavelength and intensity axes. Statistical information

was acquired, and both the preprocessed and raw datasets were subjected to principal component analysis (PCA). The machine learning models were built using two distinct train-test splits: 70% training - 30% test and 80% training - 20% test. Cross-validation was employed to keep the models from being overfit, hence the sample size is small. Both splits' machine learning outcomes from preprocessed and raw datasets were compared. This is the first time that all supervised machine learning classification algorithms, including Decision Trees, Discriminant, Nave Bayes, Support Vector Machines (SVM), k-NN (k-Nearest Neighbor), Ensemble Learning, and Neural Network algorithms, have been applied to LIBS datasets of both paracetamol-based pharmaceutical samples and 8 different mineral samples, as well as their preprocessed and raw datasets, to investigate the effect of preprocessing.

Keywords: Machine Learning, Laser-Induced Breakdown Spectroscopy, Medicines, Minerals, Principal Component Analysis, Preprocessing

ÖZET

LAZER KAYNAKLI KIRILMA SPEKTROSKOPİSİYLE FARMASÖTİK VE MİNERAL NUMUNELERİ ÜZERİNDE PCA KOMBİNE MAKİNE ÖĞRENME TEKNİKLERİNE ÖN İŞLEME YAPILMASININ PERFORMANS DEĞERLENDİRMESİ

YAZICI, Göktuğ
Y. Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Reşat Özgür DORUK

Ocak 2023, 82 sayfa

Lazerle indüklenen kırılma spektroskopisi (LIBS), malzeme tanımlama ve analiz için kullanılan hızlı bir optik atomik emisyon spektroskopisidir. Yerinde analiz, titiz numune işlemenin kaldırılması ve değerlendirilmekte olan madde için mikro yıkıcı özelliklerin avantajlarına sahiptir. LIBS, malzemeyi belirli bir eşiğe uyarmak için kısa lazer ışını patlamaları kullanır ve bu plazma oluşumuyla sonuçlanır. Dalga boyu değeri ve yoğunluk genliğini içeren plazma özellikleri, deneyin malzemesi ve çevresinden etkilenir. Bu çalışmada LIBS kullanılarak ilaç ve mineral örneklerinin spektrum profilleri elde edilmiştir. Farmasötik numunelerin toplanması, her iki parasetamol bazlı ilacın, Aferin ve Parafon'un iki farklı konsantrasyonundan oluşur. Alüminyum (Al), Bizmut (Bi), Bakır (Cu), Demir (Fe), Manganez (Mn), Nikel-Alüminyum (NiAl), Kalay (Sn), Çinko (Zn) mineral verisetindeki numunelerdir. Numunelerin spektrum verileri, eksik değerlerin şekli koruyan parçalı kübik spline enterpolasyonu ile değiştirilmesi, çeyreklere dayalı aykırı değerlerin doldurulması, gürültüyü gidermek için spektrumların yumuşatılması ve hem dalga boyu hem de

yoğunluk eksenlerinin normalleştirilmesiyle veri ön işleme yöntemlerine tabi tutulmuştur. İstatistiksel bilgiler elde edilmiş, ve hem önceden işlenmiş hem de ham veri kümeleri temel bileşen analizine (PCA) tabi tutulmuştur. Makine öğrenimi modelleri, iki farklı eğitim testi bölümü kullanılarak oluşturulmuştur: %70 eğitim - %30 test ve %80 eğitim - %20 test. Modellerin aşırı uyumlanmasını önlemek için çapraz doğrulama kullanılmış olup, bu nedenle örnek boyutu minimumdur. Her iki bölümün de önceden işlenmiş ve ham veri kümelerinden elde edilen makine öğrenimi sonuçları karşılaştırılmıştır. Karar Ağaçları, Diskriminant, Naïve Bayes, Destek Vektör Makineleri (SVM), k-NN(k-En Yakın Komşu) Topluluk Öğrenmesi ve Sinir Ağı algoritmalarından oluşan; hem parasetamol bazlı farmasötik numunelerin hem de 8 farklı mineral numunelerin LIBS veri setlerine, ve bunların hem ön işlemeye tabi tutulmuş hem de ham veri setlerine, ön işlemenin etkisini gözlemlemek için uygulandığı ilk çalışmadır.

Anahtar Kelimeler: Makine Öğrenimi, Lazer Kaynaklı Kırılma Spektroskopisi, İlaçlar, Mineraller, Temel Bileşen Analizi, Ön İşleme

DEDICATED TO

To My Dear Parents

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Dr. Reşat Özgür DORUK and Prof. Dr. Kemal Efe ESELLER, for guiding me throughout the thesis work.

Also, I would like to thank Vadi Su YILMAZ, for helping me with this thesis and kindly answering my questions.

I would also like to thank the members of the thesis defense jury.

Finally, I would like to thank my parents and my colleague Mehmet Çağdaş ÜSTÜN, for supporting and motivating me.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
k-NN / KNN	K-Nearest Neighbors
FWHM	Full Width at Half Maximum
LDA	Linear Discriminant Analysis
LIBS	Laser Induced Breakdown Spectroscopy
NaN	Not A Number
Nd:YAG	Neodymium-Doped Yttrium Aluminum Garnet; Nd:Y3Al5O12
NN	Neural Networks
PC	Principal Component
PCA	Principal Component Analysis
PLS	Partial Least Square
RSD	Relative Standard Deviation
SVM	Support Vector Machine
TDIDT	Top-Down Induction Of Decision Trees

CHAPTER 1

INTRODUCTION

LIBS (laser-induced breakdown spectroscopy) is a rapid molecular testing method that creates a micro-plasma on the sample surface using a short laser pulse [2, 5-6, 10-11, 14-18,21]. This analytical methodology offers numerous notable benefits over other elemental analysis techniques, including [2, 5-6, 10-11, 14-18, 21]:

a) A measurement experience that does not need sample preparation
b) A single spot analysis requires a relatively short measuring period (usually a few seconds)
c) A broad range of elements are addressed, including those that are lighter like H, Be, Li, C, N, O, Na, and Mg
d) A range of sampling techniques, such as a rapid raster of the sample surface and depth profiling
e) Examination of thin samples without regard for the impact of the substrate

Furthermore, its applications have expanded to encompass metallurgy, mining, environmental, and medicinal research [18-19, 21]. The pharmaceutical industry has recently come to understand the value of LIBS technology for applications such as continuous monitoring and product analysis [1]. The reading of spectrum data is the basic principle of LIBS analysis. The data provides intensity peak points in the horizontal axis that correlate to certain wavelength values. Peak values are caused by certain components in the samples. Analyzer analyses spectrum data by determining which wavelengths have the highest peak points. The user chooses the contents and concludes the experiment by determining the sample itself using chemometric methods [25].

Machine learning algorithms in material analysis are still in their inchoate phases, but they yield promising findings that challenge established approaches such as chemometrics. LIBS and PCA coupled machine learning cooperation may be used to analyze soil lead concentrations, Groundwater Fe and Cu percentages, minor steel components, gold element concentration in ores, and classification of herbal medicinal

plants [7, 23, 25–26]. The preceding research benefitted from principal component analysis (PCA) to select the significant wavelength components and preprocessing to remove the noisy components. Prior to using PCA, the aforementioned experiments used normalization, outlier reduction, and smoothing. Nevertheless, no study has been conducted to explore the influence of preprocessing on the categorization of Aferin and Parafon samples, as well as eight distinct minerals, using PCA integrated machine learning approaches. The current study investigates the efficacy of PCA combined supervised machine learning classification algorithms with preprocessing for classifying Aferin and Parafon medicines, as well as their two concentrations; and for classifying different minerals, including Aluminum (Al), Bizmut (Bi), Copper (Cu), Iron (Fe), Manganese (Mn), Nickel-Aluminum (NiAl), Tin (Sn), and Zinc (Zn).

CHAPTER 2

LIBS THEORY

Lasers are the most often used instrument for excitation, ionization, and other spectroscopic studies because they offer a distinctive light source for atomic and molecular spectroscopy (coherent, monochromatic, multimode form, directed, single mode or energetic etc) [27].

A significant number of innovative test-related procedures and applications were made possible with the help of the laser. Lasers may be used to detect a variety of innovative optical changes, and scientists can gather a significant amount of spectroscopic data pertinent to the infrared spectral region. Scientists examined strong laser beams and made several non-linear helpful spectroscopic discoveries on the Raman spectra of all types of materials via spectra. Doppler broadening could be avoided because to laser technology, which also produced resolutions close to that of natural lines. Laser spectroscopy is now more intriguing than ever because to these developments toward new experimental possibilities.

Therefore, laser-induced breakdown spectroscopy (LIBS) started to carry importance as a non-harmful material analysis method. In the following chapters, fundamentals, usage areas, advantages, and disadvantages of LIBS will be explained.

2.1. Fundamentals of LIBS

A rather intense laser pulse is used as the excitation source in LIBS, a kind of nuclear emission spectroscopy [27, 28]. A focused laser creates a plasma that excites and atomizes materials. Only when the focused laser exceeds a specific optical breakdown threshold, which varies depending on the surroundings and the target material, does plasma form [29, 30].

The electron is accelerated by inverse Bremsstrahlung after multiphoton or tunnel ionization and may interact with nearby molecules to produce more electrons. Avalanche or cascade ionization may result from the newly ionized electrons being accelerated if the pulse length is long enough [31]. A breakdown occurs at a specific

level of electron density, leading to the formation of a high density plasma with no trace of the laser pulse. As a result, the following is necessary for a pulse in thick material to be short: If the avalanche ionization threshold is not reached during the interaction, the pulse is considered short. On the appearance, this definition could seem excessively narrow. Fortunately, the precisely balanced nature of the pulses in thick medium prevents the threshold from being easily reached [32]. Intensity clamping [32] by the onset of filamentation throughout the dispersion of powerful laser pulses in dense material is the process in charge of maintaining the equilibrium.

A possible breakthrough in LIBS is the use of a short laser pulse as a spectroscopic source [33]. This technique involves concentrating ultrafast laser pulses in a gas to create a plasma column. Self-luminous plasma is significantly superior in terms of low level of continuum and decreased line broadening. This is because brief laser pulses cause a reduction in plasma density, which limits the power of the pulse in the contact zone and prevents further multiphoton/tunnel ionization of the gas [34, 35]. In a single, neutral atomic species optically thin plasma in local thermal equilibrium, the emission rate density of photons emitted by a transition from level i to level j is [36]:

$$I_{ij}(\lambda) = \frac{1}{4\pi} n_0 A_{ij} \frac{g_i \exp^{-E_i/k_B T}}{U(T)} I(\lambda) \quad (1)$$

I_{ij} is the emission rate density of photons (in m^{-3} sr^{-1} s^{-1})

no is the transition probability between level i and level j (in s^{-1})

Aij is the number of neutral atoms in the plasma (in m^{-3})

gi is the degeneracy of the upper level i (2J+1)

U(T) is the partition function (in s^{-1})

Ei is the energy level of the upper level i (in eV)

kB is the Boltzmann constant (in eV/K)

T is the temperature (in K)

$I(\lambda)$ is the line profile such that $\int_{-\infty}^{\infty} I(\lambda) d\lambda = 1$

λ is the wavelength (in nm)

The partition function *U(T)* is the statistical occupation fraction of every level *k* of the atomic species [37]:

$$U(T) = \sum_j g_j \exp^{-E_j/k_B T} \quad (2)$$

2.2. LIBS Applications

Due to its capacity to handle materials in solid, liquid, and gaseous forms (including aerosols), which may or may not be conductive, LIBS has been utilized for both qualitative and quantitative purposes in a range of matrices of relevance to numerous fields [38]. Some of these applications are exclusive to LIBS and make advantage of the method's inborn qualities, such as the capacity to do microanalysis and distant analysis and its quasi-non-destructive nature [38].

2.2.1. Metallurgical Applications: Since material pre-treatment is not necessary when using LIBS, it is a time and money-saving method that holds great promise for alloy study. Although brass, gold, and aluminum alloys have all been studied using the LIBS technique, steel is the most intriguing alloy [38].

2.2.2. Environmental Applications: The capacity to conduct in-situ field studies has sparked interest in using LIBS for environmental evaluations [38]. The main problem faced by scientists who seek to use this method to environmental studies is interference brought on by matrix effects. The situation has improved, nonetheless, thanks to the adoption of an internal standard and several other calibration techniques [38]. Most investigations in the literature are the result of assessments of the soil and water [38].

2.2.3. Archeological and Artistic Applications: The subject of cultural and historical conservation and restoration is becoming more and more in need of non-destructive analytical methodologies [38]. The evaluation of the makeup of painted surfaces, including the determination of the pigments and binders needed to date and restore them (such as oil paintings, ceramics, and frescoes), is crucial [38]. Similar to how accurate technical type identification of historical alloys and understanding of environmental alteration are essential for historical recording and preservation [38]. The advantages of LIBS over other quantitative methods for elemental analyses used to study historical discoveries are mainly related to its low invasiveness, ability to conduct in-situ measurements, good spatial discrimination, speed, and ability for direct analysis without sample treatment [38]. It is also important to note that the LIBS

technique may be used in combination with cleaning procedures, which are usually carried out with lasers, to track the process by concurrently gathering data on the makeup of the objects being cleaned while avoiding overcleaning [38]. Recently, research on LIBS's use in art and archaeology has been conducted [38]. However, the number of research on this topic has risen significantly in last several years [38].

2.2.4. Pharmaceutical Applications: LIBS's capabilities for quick analysis of multi-component pharmaceutical tablets were described [38]. According to the study, changing the plasma environment (for example, with helium flow over the tablet surface) or using internal standardization (with carbon line or plasma continuum emission) can significantly improve the analytical performance of LIBS, specifically sensitivity, linearity, precision, and matrix effect reduction [38]. Furthermore, the capabilities of LIBS for the direct and quick examination of liquid medicinal formulations were proven [38]. As a model compound for this example, sodium chloride in solution was used [38]. To create a gaseous plasma from a liquid sample, a pulsed Nd:YAG laser (1064 nm, 10 ns) was utilized [38]. The results indicated that for isotonic solutions, the RSD of a 50-second measurement (average of 50 laser shots at 1 shot s-1) utilizing surface analysis of a flowing solution was roughly 0.5% [38].

2.2.5. Biological and Medical Applications: It was stated that a quantitative LIBS analysis was done to investigate the presence of trace minerals in teeth [39-41]. The research focused on teeth from various age groups [39-41]. The samples revealed several components that are ordinarily found in trace levels in biological material. The quantities of aluminum discovered in the teeth under analysis were determined to be most likely due to the use of whitening toothpaste and the presence of fillings [39-41].

CHAPTER 3

MACHINE LEARNING THEORY

3.1. Decision Trees

Decision tree learning is a popular data mining method [42]. The goal is to build a model that predicts the value of a target variable based on a number of input variables. A decision tree is a straightforward representation for categorizing examples. Assume that all of the input features have finite discrete domains and that there is a single target feature called "classification" for this section. Each element of the classification domain is referred to as a class. A decision tree, also known as a classification tree, is a tree in which each internal (non-leaf) node has an input feature labeled. A tree is constructed by dividing the source set, which serves as the tree's root node, into subsets, which serve as the tree's successor children [43]. The splitting is based on a set of classification-based splitting rules [43]. This process is repeated recursively on each derived subset, which is known as recursive partitioning [43]. When the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions, the recursion is complete [43]. This aggressive approach, known as top-down induction of decision trees (TDIDT) [44], is by far the most used method for inferring decision trees from data. Regression tree analysis and classification tree analysis are the two decision tree types utilized in data mining [45, 46]. The former is employed when the class (discrete) to which the data belongs is the anticipated outcome [45, 46]. The latter is utilized when the outcome being forecasted is a real figure, such as the cost of a home or the duration of a patient's medical care [46]. Both of the aforementioned techniques are referred to as classification and regression tree (CART) analysis and were initially presented by Breiman et al. in 1984 [46]. Regression and classification trees have certain similarities and distinctions, such as how they choose where to partition data [46].

3.2. Discriminant Analysis

Discriminant analysis is a classification method based on the assumption that different classes generate data using different Gaussian distributions [48]. The fitting

function estimates the parameters of a Gaussian distribution for each class to create a discriminant classifier [48]. The trained classifier finds the class with the lowest misclassification cost to predict the classes of new data [48]. Furthermore, a type of discriminant analysis known as linear discriminant analysis is also known as the Fisher discriminant, named after its inventor, Sir R. A. Fisher [48].

3.3. Naïve Bayes

Naïve Bayes classifiers are a type of simple "probabilistic classifier" that uses Bayes' theorem with strong (naive) independence assumptions between features (see Bayes classifier). They are among the most basic Bayesian network models [49], but when combined with kernel density estimation, they can achieve high levels of accuracy [50]. Naïve Bayes classifiers are highly scalable, with parameters that are proportional to the number of variables (features/predictors) in a learning problem. Maximum-likelihood training can be done in linear time by evaluating a closed-form expression [51], as opposed to the expensive iterative approximation used for many other types of classifiers.

3.4. Support Vector Machines

Support vector machines (SVMs, also known as support vector networks [47]) are supervised learning models with associated learning algorithms that examine data for classification and regression analysis in machine learning [47]. Vladimir Vapnik and colleagues created it at AT&T Bell Laboratories [47]. SVMs are among the most robust prediction approaches, as they are based on statistical learning frameworks or the VC theory established by Vapnik and colleagues [47]. Given a series of training examples, each of which has been labeled as belonging to one of two categories, an SVM training method constructs a model that assigns subsequent instances to one of the two categories, resulting in a non-probabilistic binary linear classifier [47].

3.5. K-Nearest Neighbors

The K-nearest neighbors algorithm (k-NN) is a non-parametric supervised machine learning technique created in 1951 by Evelyn Fix and Joseph Hodges [48-52] and later modified by Thomas Cover [53]. Its applications include classification and regression. The input in both cases consists of the k closest training examples in a

data collection. k-NN is a classification method in which the function is only approximated locally and all computation is postponed until after the function has been evaluated [52, 53]. Because this method depends on distance for classification, if the features represent distinct physical units or have wildly different scales, normalizing the training data can significantly increase its accuracy [54, 55]. An effective strategy for both classification and regression is to apply weights to the contributions of the neighbors, such that the closer neighbors contribute more to the average than the further ones [54, 55].

3.6. Ensemble Learning

Ensemble approaches in statistics and machine learning employ many learning algorithms to achieve greater prediction performance than each of the constituent learning algorithms alone [56-58]. In contrast to a statistical ensemble in statistical mechanics, which is normally unlimited, a machine learning ensemble consists only of a specific finite number of different models, but typically allows for considerably more flexible structure to exist among those alternatives [56-58]. Empirical evidence suggests that ensembles produce better outcomes when the models are diverse [59, 60]. As a result, several ensemble approaches strive to enhance variety among the models they combine [61, 62]. Despite its appearance as counterintuitive, more random algorithms (such as random decision trees) can create a stronger ensemble than more purposeful algorithms (such as entropy-reducing decision trees) [63]. Nevertheless, it has been demonstrated that utilizing a range of powerful learning algorithms is more beneficial than strategies that aim to dumb down the models in order to foster diversity [64]. In the training stage of the model, it is feasible to boost variety by employing correlation for regression tasks [65] or information metrics such as cross entropy for classification tasks [66].

3.7. Artificial Neural Network

Artificial neural networks (ANNs) are computational models which are modeled after the biological neural networks that make up animal brains [67]. An ANN is built from a network of linked units or nodes known as artificial neurons, which are roughly modeled after the neurons in the human brain [67]. Each link, like synapses in a biological brain, has the ability to relay a signal to other neurons [67]. An artificial

neuron receives impulses, analyses them, and can signal neurons to which it is attached [68]. The "signal" at each link is a real number, and each neuron's output is generated by some non-linear function of the sum of its inputs [68]. The connections are referred to as edges [68]. The weight of neurons and edges is often adjusted as learning progresses [68].

CHAPTER 4

PREPROCESSING THEORY

Data preprocessing is a process in the data mining process that involves manipulating or removing data before it is utilized to assure or improve performance [69]. The adage "garbage in, trash out" applies especially to data mining and machine learning applications. Data collection methods are frequently poorly managed, resulting in out-of-range numbers (for example, Income: 100), impossible data combinations (for example, Sex: Male, Pregnant: Yes), and missing information, among other things.

Data analysis that has not been thoroughly checked for such issues might give misleading results. Thus, before doing any analysis, the representation and quality of data must come first [70]. Data preparation is frequently the most critical stage of a machine learning project, particularly in computational biology [71].

When there is a lot of irrelevant and redundant information or noisy and inaccurate data, knowledge discovery becomes more challenging during the training phase [71]. Data preparation and filtering can consume a significant amount of processing time [71]. Cleaning, instance selection, normalization, one hot encoding, transformation, feature extraction and selection, and so on are examples of data preparation [71]. The final training set is the result of data preprocessing [71].

Data pretreatment may influence how final data processing results are understood [72]. This should be carefully studied when the interpretation of the results is critical, such as in multivariate chemical data processing (chemometrics) [72]. Data preprocessing steps can be gathered under four groups and they are:

4.1. Data cleansing: It refers to finding incomplete, erroneous, inaccurate, or unnecessary sections of the data and then changing, amending, or eliminating the filthy or coarse data [73]. Data cleaning can be done interactively using data wrangling tools, or in batches using scripting or a data quality firewall.

4.2. Data editing: It is described as the process of reviewing and adjusting gathered survey data. By correcting conflicting data using various methods [74], data editing helps develop standards that will eliminate possible bias and assure consistent

estimates, leading to a clear analysis of the data set. The goal is to ensure the accuracy of the data acquired [75]. Data editing can be done manually, using a computer, or a mix of the two [76].

4.3. Data reduction: It is the process of converting empirically or experimentally obtained numerical or alphabetical digital information into a rectified, ordered, and simpler form. Data reduction can serve two purposes: it can lower the quantity of data records by removing erroneous data or it can create summary data and statistics at various aggregation levels for various applications [77]. Thus, data reduction may be carried out on dimension(s) and volume of data.

For reducing the dimension of data, dimensionality reduction is realized. If data gets increasingly scarce as dimensionality grows, concentration and separation between points become less relevant. Dimensionality reduction reduces noise in the data and facilitates viewing. Wavelet transform is a method of dimensionality reduction in which data is changed to retain relative distance between objects at different levels of resolution and is commonly used for picture compression [78].

For reducing he volume of data, numerosity reduction is realized. It minimizes data volume by using alternate, smaller data representations. There are two types of number reduction methods: parametric and non-parametric [78]. Non-parametric approaches, such as histograms, clustering, and sampling, do not require models and instead assume that the data matches some model, estimate model parameters, save just the parameters, and discard the data [78].

4.4. Data wrangling: Data wrangling, also known as data munging, is the act of changing and mapping data from one "raw" data type to another in order to make it more suitable and valuable for a range of downstream uses such as analytics [79]. The purpose of data wrangling is to ensure that data is of high quality and usable [79]. Data analysts frequently spend most of their time wrangling data rather to really analyzing the data [79]. Data wrangling may include more munging, data visualization, data aggregation, training a statistical model, and a variety of other possible applications [79]. Data wrangling typically consists of a series of general steps that include extracting raw data from a data source, "munging" the raw data (e.g. sorting) or parsing

the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use [79].

CHAPTER 5

LITERATURE SURVEY

In literature, there are applications which combine machine learning and LIBS. These applications vary from analysis of trace of certain elements in ores from the classification of medicinal herbs. Here are some examples of the applications.

5.1. Laser-Induced Breakdown Spectroscopy and Principal Component Analysis for the Classification of Spectra from Gold-Bearing Ores [7]: In this research, laser-induced breakdown spectroscopy (LIBS) and principal component analysis (PCA) were used to classify LIBS spectra from gold ores formed as pressed granules from powdered bulk material [7]. The aim of the research was to capture gold (Au) emission lines. Principal component analysis (PCA) was used to analyze two spectral ranges of 21 nm and 0.15 nm in width [7]. After PCA, only three principal components (PCs) remained as significant variables [7]. Outlier elimination was carried out. As a result, PCA successfully identified Au emission lines [7].

5.2. Comparison of different calibration techniques of laser induced breakdown spectroscopy in bakery products: on NaCl measurement [4]: LIBS was employed in this work to demonstrate its use as a routine analysis for Na values in bakery items [4]. To examine alternative calibration approaches, a set of standard bread samples with varying amounts of NaCl (0.025%-3.5%) were created [4]. Calibration methodologies included standard calibration curve (SCC), artificial neural network (ANN), and partial least square (PLS) [4]. PLS was shown to be more efficient for predicting Na contents in bakery items, with a coefficient of determination value rise from 0.961 to 0.999 for standard bread samples and from 0.788 to 0.943 for commercial products [4].

5.3. Classification of Chinese Herbal Medicine by Laser Induced Breakdown Spectroscopy with Principal Component Analysis and Artificial Neural Network [23]: In this study, the roots of Angelica pubescens, Codonopsis pilosula, and Ligusticum wallichii from various locations of China are examined; and identified using laser-induced breakdown spectroscopy (LIBS) coupled with principal

component analysis (PCA), and artificial neural networks (ANN) [23]. Angelica pubescens, Codonopsis pilosula, and Ligusticum wallichii classification accuracy rates were 99.89%, 95.83%, and 99.85%, respectively [23]. The results surpassed both LDA and SVM as machine learning methods [23].

5.4. Estimation of the Fe and Cu Contents of the Surface Water in the Ebinur Lake Basin Based on LIBS and a Machine Learning Algorithm [25]: Water samples were taken from the Ebinur Lake Basin for this investigation [25]. LIBS was used to extract the distinctive peaks of iron (Fe) and copper (Cu) from water samples [25]. The LIBS curve had a number of peaks from each sample [25]. The intensities of the Fe and Cu characteristic lines, transition probabilities, and high signal-to-background ratio (S/B) were eventually used to identify the Fe and Cu characteristic analytical lines [25]. To improve the data, the preprocessing techniques of aberrant value elimination, mean processing, baseline removal, and wavelet analysis were used [25]. Back propagation (BP) neural network was used in the machine learning section, and its results were compared to those of other machine learning methods [25].

CHAPTER 6

EXPERIMENTAL SETUP

LIBS experiments were performed by using a LIBS whole system (Applied Spectra, J200 LA), a Quantel-Big Sky Nd: YAG-laser (Bozeman, MT, USA), HR2000 Ocean Optics Spectrograph (Dunedin, FL, USA) and Stanford Research System Delay Generator SRS DG535 (Cleveland, OH, USA). The experimental setup for LIBS is shown in Figure 1. Q-switched Nd: YAG laser (Quantel, Centurion) was used as the excitation source and it operates at 532 nm with maximum energy of 18 mJ per pulse approximately 9 ns (FWHM) pulse duration with an adjustable laser repetition range between 1 and 100 Hz. However, 1 Hz is only used for the experiment. The beam diameter at the exit was measured as 3 mm of which divergence is 5 mrad. A lens with a 50 mm focal length was used to focus the beam size onto the medicine samples. A pickup lens of 50 mm diameter was used to collect the emitted plasma which was aligned at approximately 90º with respect to the laser beam. Then, the fiber tip of the spectrometer was used to collect the plasma. The pickup lens was placed away about 15 cm to the focal point of the laser beam. The resolution of the spectrograph is approximately 0.5 nm in the 200–1100 nm range. The measurements were taken under the ambient conditions exposing the atmosphere. Samples were measured by the LIBS technique in triplicate, scanning five different locations and four excitations per location. Figure 6.1 shows the experimental setup for LIBS.

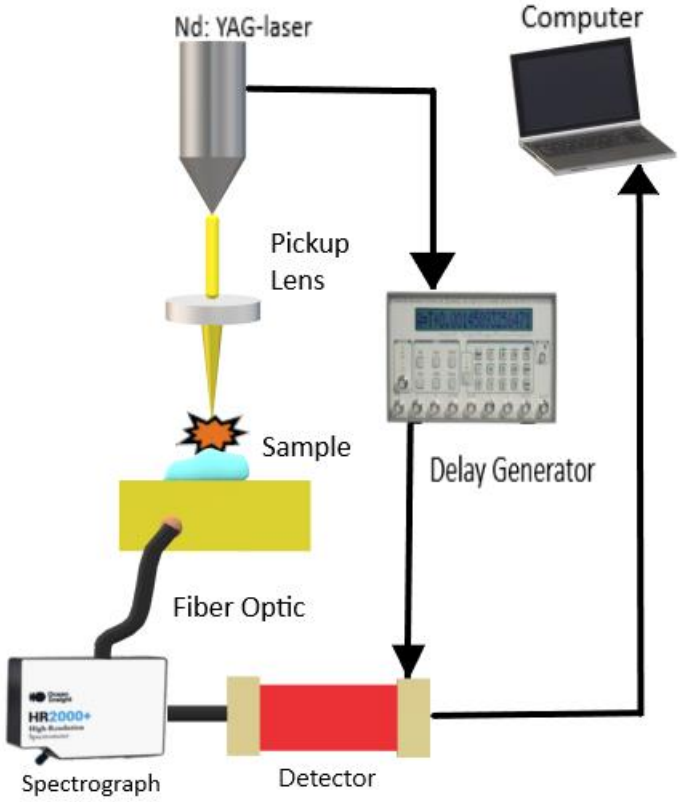

Figure 6.1 LIBS Experimental Setup

CHAPTER 7

MATERIAL, METHOD AND DATA ANALYSIS

Randomly distributed pharmaceutics and mineral samples were excited with Q-switched Nd: YAG laser then; their spectroscopic data, amplitude and wavelength values, were obtained. From these values, pharmaceutic and mineral datasets were created. The dataset used in the experiment of pharmaceutics contained two different medicines, each with two different concentrations while the dataset of minerals contained 8 different minerals. The medicines were Aferin and Parafon and their respective concentrations were 500 mg, 650 mg for Aferin and 300 mg, 500 mg for Parafon. The minerals were Aluminum (Al), Bizmut (Bi), Copper (Cu), Iron (Fe), Manganese (Mn), Nickel-Aluminum (NiAl), Tin (Sn), and Zinc (Zn). The pharmaceutics dataset contained 80 observations, 20 for each 4 medicine classes. The minerals dataset contained 24 observations, 3 for each 8 mineral classes. The classification operation was performed with machine learning tools and algorithms. The LIBS data of both datasets contained the wavelength and intensity measurements. In pharmaceutics dataset, the lowest and highest intensity measurements were -204 and 64644, respectively. In minerals dataset, the lowest and highest intensity measurements were -286 and 40215, respectively. MATLAB Programme (2021 Academic Version) was used for data analysis and machine learning steps. Python programme, with Numpy and the Pandas libraries, was used to create group colored scatter plot graphics based on the classes of datasets to visualize the first 3 PCs. The wavelength-intensity graphic of the medicines and minerals datasets are shown in Figure 7.1 and Figure 7.2, respectively.

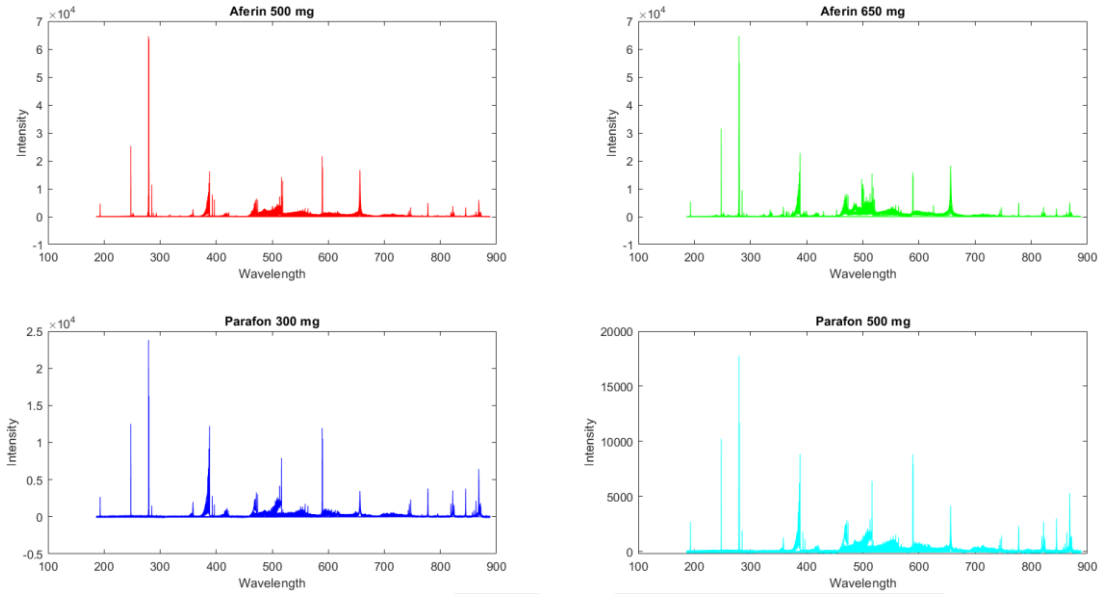

Figure 7.1 Wavelength-Intensity Graphic of Two Medicines with Two Different Concentrations

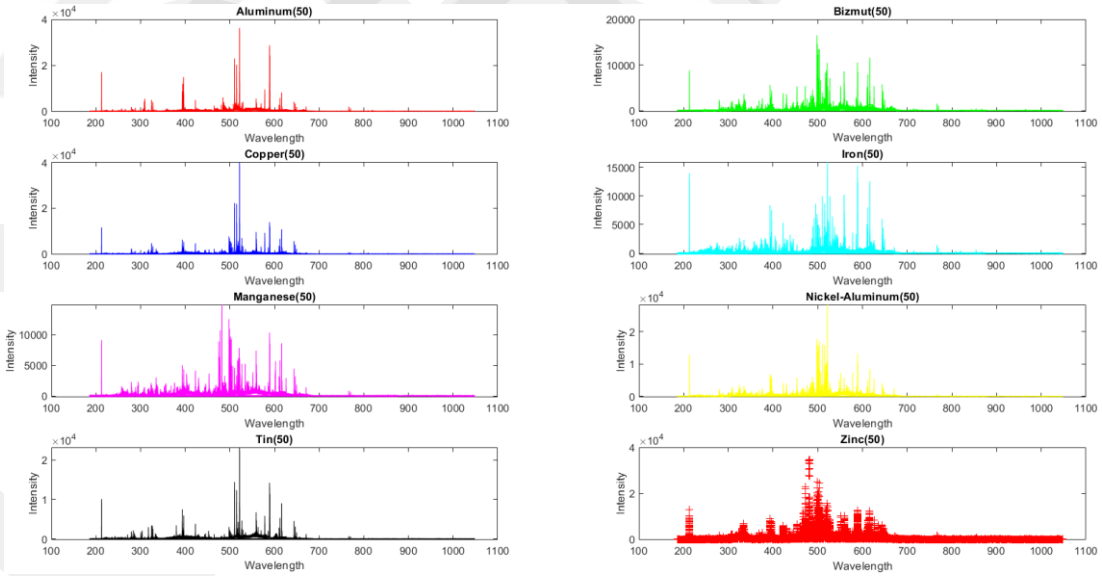

Figure 7.2 Wavelength-Intensity Graphic of 8 Different Minerals

Pharmaceutics dataset did not contain either missing or infinite values, therefore handling missing values process was skipped for them. Although, minerals dataset contained 37 NaN (not a number) values, which can be translated as missing and infinite values. Thus, these NaN values handled with shape-preserving piecewise cubic spline interpolation preprocessing based on quartiles. Furthermore, they contained outlying values which resided above 75% and under 25% quartiles. Thus, these outliers were replaced with same interpolation method. After these two preprocessing operations, the LIBS datasets were not still eligible for modelling due to the noise. As Tognoni stated, the aforementioned noise can be categorized under source noise, shot noise, detector noise and instrumental (thermal) drift [23-24]. Source noise is caused by fluctuations in the laser-sample or laser-plasma interaction; shot noise is caused by the number of photons arriving on the detector; detector noise is caused by the number of photons arriving on the detector; and instrumental (thermal) drift is caused by laser energy increase originated from optical and electronic components' warming up [23-24]. Therefore, the smoothing of LIBS spectra was applied. For the smoothing method, Savitzky-Golay filter, which smooths according to a quadratic polynomial that is fitted over each window of the data, was applied due to its effectiveness for rapidly changing values.

Next, a normalization procedure was applied for both the intensity and the wavelength values for the datasets. In machine learning applications, variable values having considerably much greater values than the other variables' values have tendency for leading to inaccurate models. Therefore, the normalization was performed for both variables, meaning intensity and wavelength values of the both datasets, by placing them in 0-1 range. The applied formula for the normalization is:

$$y_i = \frac{x_i - min(x)}{max(x) - min(x)} \tag{3}$$

Train and test datasets were created with %70-30% and %80-20% for both the raw and preprocessed datasets, and they were saved to prevent the randomness effect. The saved datasets used in the PCA and modeling steps later. During PCA, each wavelength was transformed into a feature. Thus, the medicine dataset transformed into an 80x10241, 80 medicine samples' intensity values as observations and 10241 wavelength values as features. The mineral dataset transformed into an 960x12289,

960 mineral samples' intensity values as observations and 12289 wavelength values as features. In the PCA step, MATLAB was used to find PCs values. But MATLAB's version in this study did not have a predefined function to plot PCs with their respective groups. Therefore, the Python program, with Numpy and the Pandas libraries, was used to create scatter plot graphics grouped by medicines and mineral classes on the train datasets. For the machine learning-based modeling step of two datasets, MATLAB's Classification Learner Application was used. First, cross-validation with 10 folds was applied to protect against overfitting. Afterward, PCA was applied and the component reduction criterion was specified with respect to explaining 95% of the variance. Finally, all applicable classifiers (decision trees, discriminant models, naïve bayes, support vector machines, nearest neighbor, ensemble learning, neural networks) with default settings were chosen to model the datasets.

CHAPTER 8

RESULTS AND DISCUSSION

The raw and preprocessed versions of the both pharmaceutics and mineral datasets were split into train and test parts with 70%-30% and 80%-20% ratios. In the PCA step, it was aimed that PCs explained 95% variance of the whole data. For the raw pharmaceutics dataset with the train-test split of 70%-30%, the first 4 of 55 principal components resulted sufficient to explain the desired variance while for its 80%-20% split version, the first 5 of 63 components were enough. However, for the preprocessed medicine data with 70%-30% and 80%-20% train-test splits, the minimum number of principal components increased to 38 of 55 and 44 of 63, respectively. For the raw minerals dataset with the train-test split of 70%-30%, the first 15 of 646 principal components resulted sufficient to explain the desired variance while for its 80%-20% split version, the first 15 of 738 components were enough. However, for the preprocessed minerals data with 70%-30% and 80%-20% train-test splits, the minimum number of principal components increased to 521 of 671 and 590 of 738, respectively. Table 8.1 and Table 8.2 show the explained percentage variance of PC values.

Table 8.1 Percentage of Explained Variances of First 15 Principal Components of Pharmaceutics Dataset

Percentage of Explained Variances for First 15 Principal Components of Pharmaceutics Dataset			
70% Train - 30% Test		80% Train - 20% Test	
Raw Data	Preprocessed Data	Raw Data	Preprocessed Data
78.382	65.746	76.394	65.834
11.152	4.210	11.295	3.949
3.163	2.457	3.603	2.685
2.460	2.213	2.897	2.175
1.249	1.529	1.818	1.399
0.801	1.447	0.953	1.337
0.557	1.123	0.598	1.008
0.427	1.001	0.364	0.913
0.314	0.848	0.316	0.784

0.241	0.809	0.298	0.742
0.181	0.752	0.180	0.698
0.122	0.738	0.123	0.686
0.088	0.696	0.111	0.638
0.074	0.657	0.092	0.593
0.064	0.639	0.076	0.570

Table 8.2 Percentage of Explained Variances of First 15 Principal Components of Minerals Dataset

Percentage of Explained Variances for First 15 Principal Components of Minerals Dataset			
70% Train - 30% Test		80% Train - 20% Test	
Raw Data	Preprocessed Data	Raw Data	Preprocessed Data
30.292	21.604	29.936	20.589
18.033	11.878	17.842	11.850
10.649	5.604	11.701	5.662
10.0216	4.062	9.742	4.111
6.496	3.567	6.626	3.983
6.260	2.445	5.894	2.461
5.027	2.139	5.622	2.292
3.526	1.745	3.418	1.732
1.528	1.204	1.471	1.287
0.894	0.973	0.844	0.990
0.677	0.854	0.629	0.814
0.543	0.7511	0.502	0.791
0.444	0.729	0.418	0.746
0.350	0.702	0.339	0.664
0.340	0.654	0.325	0.626

It was observed that preprocessing caused a great increase in the minimum number of PCs in both datasets. Since PCA is highly sensitive to data scaling, and the raw datasets' features contained highly varying values compared to each other (one feature varies between 0-50 while the other one varies between 200-250, eg.), it was expected from PCA to assign more weights to the components related with the features oscillating between greater numbers. After the preprocessing, especially due to the effect of normalization, every feature value oscillated between 0-1. Thus, it was

ensured that all features were assigned with equally importance. After the PCA step, all classification machine learning models of MATLAB with default parameters were applied to both raw and preprocessed pharmaceutics and minerals datasets with both train-test splits. For both raw and preprocessed pharmaceutics datasets, the train-test split of 70%-30% resulted in 56 samples for training and 24 samples for testing, while the train-test split of 80%-20% resulted in 64 samples for training and 16 samples for testing. For both raw and preprocessed minerals datasets, the train-test split of 70%-30% resulted in 672 samples for training and 288 samples for testing, while the train-test split of 80%-20% resulted in 768 samples for training and 292 samples for testing. Training datasets of medicines and minerals samples were then applied PCA, and their respected figures for first 3 PCs are shown in Figures 8.1 to 8.4 for medicines, and Figures 8.5 to 8.8 for minerals.

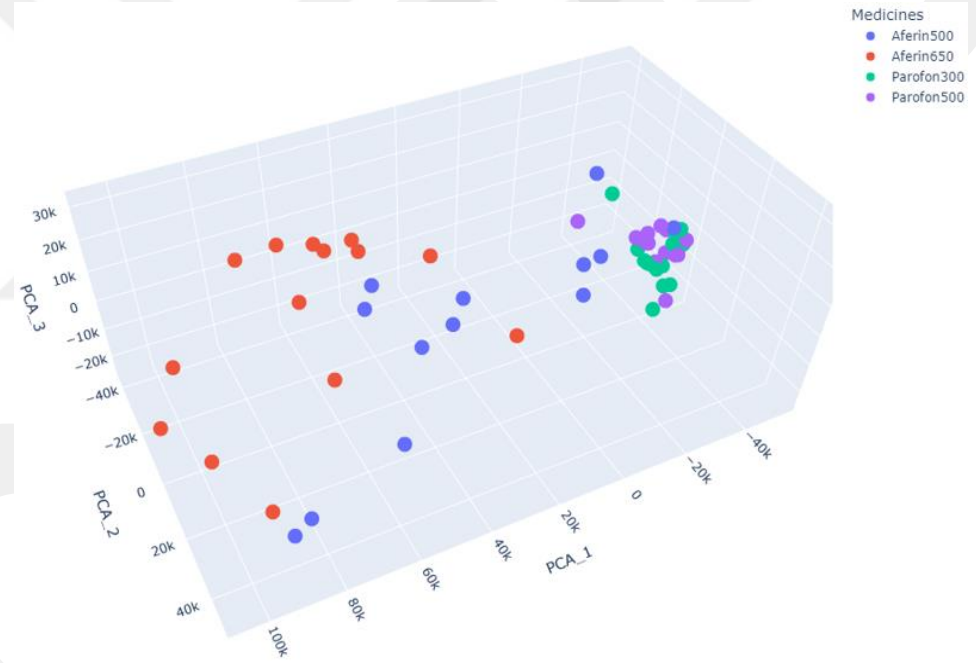

Figure 8.1 Scatterplot of Medicines Raw Train Dataset's First 3 PCs with 70%-30% Train-Test Split

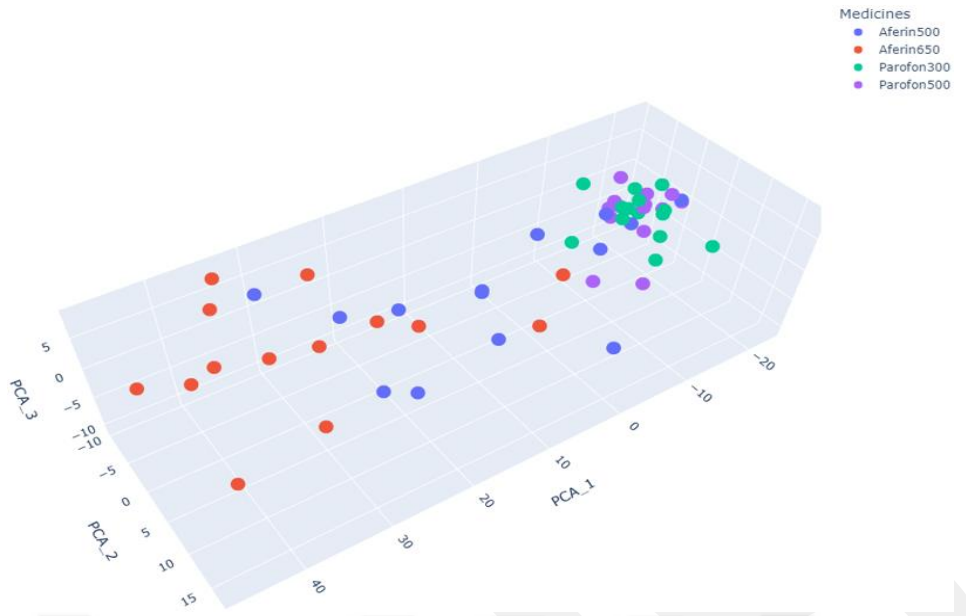

Figure 8.2 Scatterplot of Medicines Preprocessed Train Dataset's First 3 PCs with 70%-30% Train-Test Split

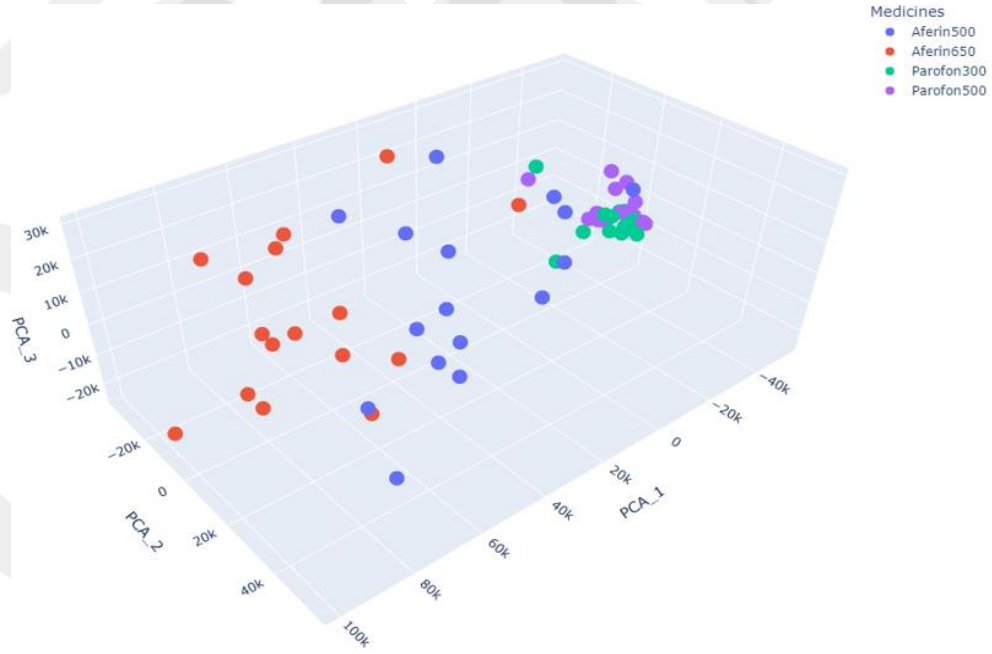

Figure 8.3 Scatterplot of Medicines Raw Train Dataset's First 3 PCs with 80%-20% Train-Test Split

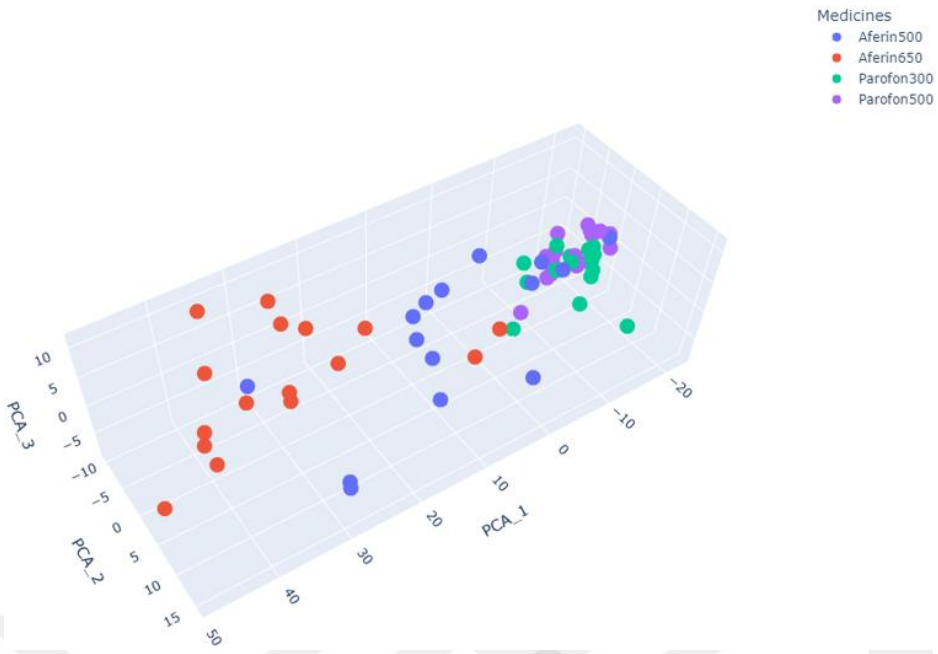

Figure 8.4 Scatterplot of Medicines Preprocessed Train Dataset's First 3 PCs with 80%-20% Train-Test Split

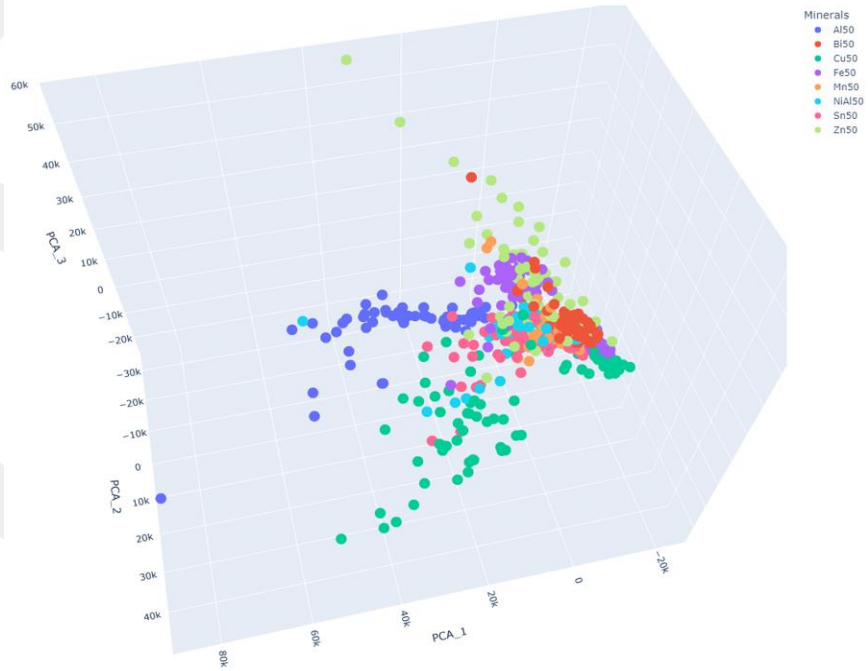

Figure 8.5 Scatterplot of Minerals Raw Train Dataset's First 3 PCs with 70%-30% Train-Test Split

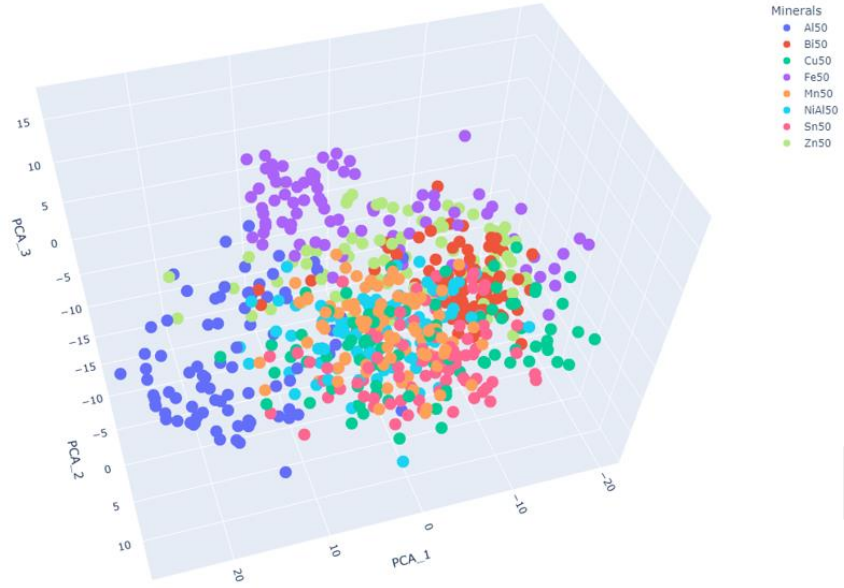

Figure 8.6 Scatterplot of Minerals Preprocessed Train Dataset's First 3 PCs with 70%-30% Train-Test Split

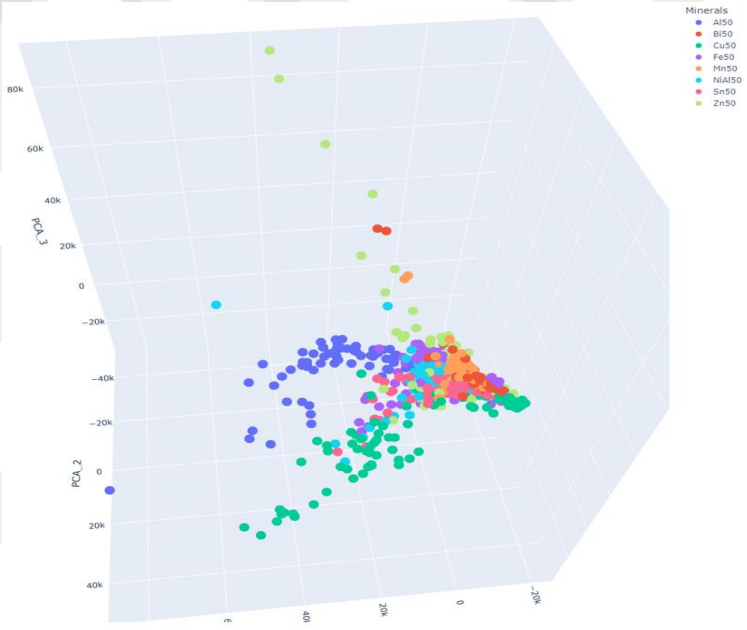

Figure 8.7 Scatterplot of Minerals Raw Train Dataset's First 3 PCs with 80%-20% Train-Test Split

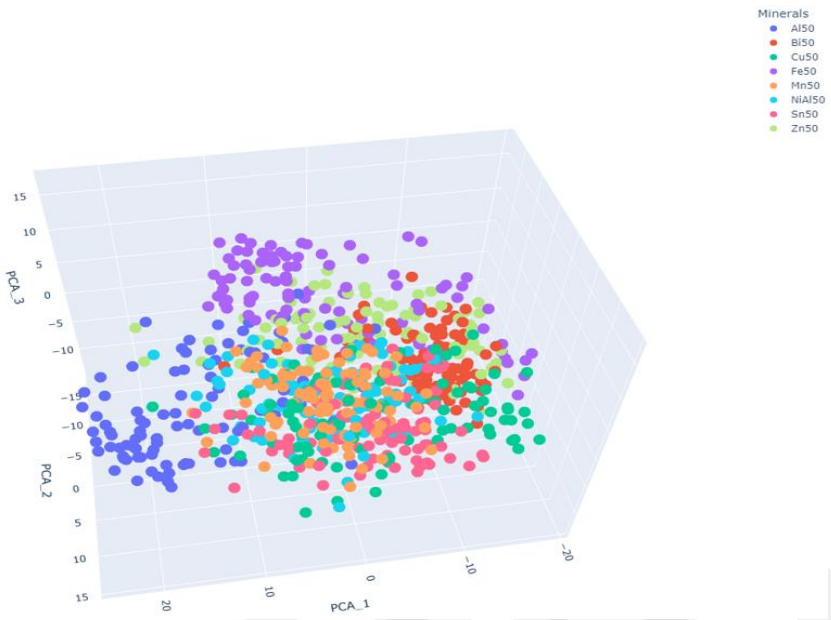

Figure 8.8 Scatterplot of Minerals Preprocessed Train Dataset's First 3 PCs with 80%-20% Train-Test Split

Both raw and processed PCA applied train datasets were modeled by the classification learners aforementioned in the Section 3. The results of raw and processed datasets were compared. Table 8.3 and Table 8.4 show results of the machine learning models for medicines and minerals datasets, respectively.

Table 8.3 Machine Learning Results of The Medicines Dataset

	70%-30% Train Test Split				80%-20% Train Test Split			
	Validation Accuracy (%)		Test Accuracy (%)		Validation Accuracy (%)		Test Accuracy (%)	
Classification Models	Raw Data	Preprocessed Data	Raw Data	Preprocessed Data	Raw Data	Preprocessed Data	Raw Data	Preprocessed Data
Fine Tree	48.2	57.1	66.7	66.7	57.8	54.7	50.0	68.8
Medium Tree	48.2	57.1	66.7	66.7	57.8	54.7	50.0	68.8
Coarse Tree	50.0	55.4	66.7	70.8	64.1	53.1	43.8	68.8

Linear Discriminant	53.6	80.4	66.7	83.3	64.1	84.4	37.5	87.5
Quadratic Discriminant	55.4	58.9	58.3	54.2	64.1	57.8	50.0	50.0
Gaussian Naïve Bayes	53.6	58.9	58.3	54.2	64.1	57.8	37.5	50.0
Kernel Naïve Bayes	51.8	62.5	41.7	70.8	67.2	60.9	37.5	68.8
Linear SVM	50.0	62.5	58.3	87.5	64.1	70.3	43.8	75.0
Quadratic SVM	57.1	67.9	58.3	83.3	71.9	65.6	68.8	75.0
Cubic SVM	64.3	57.1	54.2	75.0	64.1	56.2	62.5	75.0
Fine Gaussian SVM	30.4	35.7	54.2	66.7	23.4	21.9	43.8	75.0
Medium Gaussian SVM	30.4	37.5	50.0	70.8	23.4	21.9	31.2	75.0
Coarse Gaussian SVM	30.4	35.7	50.0	70.8	23.4	23.4	31.2	75.0
Fine KNN	58.9	32.1	45.8	33.3	54.7	31.2	50.0	37.5
Medium KNN	57.1	32.1	54.2	37.5	62.5	31.2	31.2	31.2
Coarse KNN	17.9	17.9	25.0	25.0	15.6	15.6	25.0	25.0
Cosine KNN	57.1	66.1	50.0	79.2	60.9	76.6	37.5	75.0
Cubic KNN	55.4	25.0	54.2	37.5	56.2	29.7	37.5	31.2
Weighted KNN	58.9	33.9	54.2	33.3	64.1	34.4	56.2	31.2
Ensemble Boosted Trees	17.9	17.9	25.0	25.0	15.6	15.6	25.0	25.0
Ensemble Bagged Trees	57.1	62.5	66.7	75.0	73.4	58.4	43.8	75.0
Ensemble Subspace Discriminant	53.6	83.9	66.7	83.3	60.9	84.4	37.5	87.5
Ensemble Subspace KNN	53.6	67.9	54.2	79.2	65.6	73.4	50.0	75.0

Ensemble RUSBoosted Trees	42.9	35.7	25.0	25.0	37.5	25.0	25.0	25.0
Narrow Neural Network	58.9	66.1	45.8	79.2	59.4	71.9	50.0	75.0
Medium Neural Network	58.9	75.0	58.3	83.3	62.5	75.0	56.2	68.8
Wide Neural Network	50.0	78.6	58.3	79.2	65.6	79.7	56.2	75.0
Bilayered Neural Network	57.1	62.5	58.3	70.8	65.6	62.5	50.0	68.8
Trilayered Neural Network	53.6	66.1	50.0	66.7	60.9	65.6	62.5	68.8

Table 8.4 Machine Learning Results of The Minerals Dataset

	70%-30% Train Test Split				80%-20% Train Test Split			
	Validation Accuracy (%)		Test Accuracy (%)		Validation Accuracy (%)		Test Accuracy (%)	
Classification Models	Raw Data	Preprocessed Data	Raw Data	Preprocessed Data	Raw Data	Preprocessed Data	Raw Data	Preprocessed Data
Fine Tree	80.8	73.2	78.8	76.7	82.7	72.3	79.2	78.6
Medium Tree	76.2	72.3	66.3	72.9	74.6	67.1	72.4	71.9
Coarse Tree	42.9	52.1	41.0	53.1	44.5	52.0	37.0	52.6
Linear Discriminant	92.0	98.4	90.6	99.0	91.8	98.7	90.6	97.9
Quadratic Discriminant	94.5	82.6	93.8	82.3	94.5	80.6	90.6	75.5
Gaussian Naïve Bayes	85.0	82.6	81.2	82.3	84.6	80.6	85.4	75.5
Kernel Naïve Bayes	87.9	55.5	82.6	48.3	86.3	52.1	83.9	47.4

Linear SVM	98.2	94.9	98.3	99.3	98.6	95.7	98.4	97.4
Quadratic SVM	98.8	99.6	99.3	99.3	98.7	99.3	99.0	99.0
Cubic SVM	98.2	99.3	98.6	99.3	99.3	99.2	99.0	99.0
Fine Gaussian SVM	86.5	87.8	85.8	89.9	87.8	86.5	88.0	82.3
Medium Gaussian SVM	46.7	57.0	70.8	99.0	33.9	37.2	71.9	97.4
Coarse Gaussian SVM	47.0	57.1	70.8	99.3	33.9	37.2	71.9	97.4
Fine KNN	92.6	47.5	91.0	47.9	93.9	47.4	92.2	43.2
Medium KNN	92.1	56.0	91.7	54.2	91.9	52.0	90.6	48.4
Coarse KNN	77.1	63.7	78.1	63.9	79.7	59.0	81.8	54.7
Cosine KNN	91.8	94.3	91.3	96.2	91.4	93.2	90.6	91.1
Cubic KNN	92.4	55.8	91.3	56.2	91.7	53.4	91.7	55.2
Weighted KNN	93.5	57.3	92.4	54.5	93.2	53.4	92.2	49.5
Ensemble Boosted Trees	83.8	82.0	83.7	87.2	87.0	82.0	86.5	85.4
Ensemble Bagged Trees	91.1	85.7	87.8	87.8	92.1	84.9	91.7	85.9
Ensemble Subspace Discriminant	91.8	98.8	91.0	99.0	91.9	98.4	91.1	97.9
Ensemble Subspace KNN	91.5	87.5	90.6	88.2	91.8	87.9	93.2	89.6
Ensemble RUSBoosted Trees	83.6	71.7	80.6	72.9	83.9	69.0	80.2	70.8
Narrow Neural Network	95.1	94.6	94.1	96.2	95.3	95.3	93.8	93.2
Medium Neural Network	95.8	98.1	93.1	98.3	95.8	98.4	93.8	99.0
Wide Neural	95.8	99.0	94.1	99.3	96.1	98.3	93.8	97.9

Network								
Bilayered Neural Network	95.4	82.6	93.1	85.1	95.3	83.5	91.7	93.2
Trilayered Neural Network	95.4	69.0	96.2	81.6	94.0	73.4	92.7	85.9

In the medicines dataset, preprocessing was found to enhance test accuracies for both train-test splits, particularly for SVM and Neural Network (NN) models. It did, however, have a considerable negative impact on a few KNN models (Fine, Medium, Cubic, and Weighted KNN), Quadratic Discriminant, Gaussian Naïve Bayes, Ensemble RUSBoosted Trees, and Trilayered NN models. The accuracy loss of KNN models can be explained by datapoint convergence owing to normalization and outlier eradication. Because normalization reduces distances between data points by scaling and outlier elimination removes extreme values, the ability of the KNN classifier to discriminate between classes becomes less accurate than previously. Furthermore, preprocessing considerably enhanced the validation accuracy of Linear Discriminant and Ensemble Subspace Discriminant models. However, due to the aforementioned effects of normalization and outlier reduction, the same preprocessing had a negative impact on KNN models. In general, preprocessing improved validation accuracies for the remaining models.

In the minerals dataset, preprocessing considerably enhanced the overall accuracies, validation and test accuracies of both of the splits, of Coarse Tree, Linear Discriminant, Ensemble Subspace Discriminant, Medium Gaussian SVM, and Coarse Gaussian SVM models. Furthermore, preprocessing made minor improvements on Medium Neural Network and Wide Neural Network models' overall accuracies. However, it caused a major negative impact on a few KNN (Fine, Medium, Coarse, Cubic, and Weighted KNN), Tree (Fine and Medium), Ensemble (Bagged Trees, RUSBoosted Trees, and Subspace KNN), and Neural Network (Bilayered and Trilayered) models. Similarly to medicines dataset's results; KNN models, and of Ensemble Subspace KNN model, accuracy losses are originated from the effects of

normalization and outlier reduction. Additionally, overall accuracy drop of Tree models, and of Ensemble Bagged and RUSBoosted Tree models, are also caused by the normalization and outlier reduction, along with working on 8 classes instead of 4 classes. Apart from this, accuracy drop of Bilayered NN, Trilayered NN, Quadratic Discriminant, and Kernel Naïve Bayes models require further investigation on the root cause.

Contrary to popular belief, validation accuracies, or train dataset accuracy, were lower than their corresponding test accuracies in this study. It should be noted that the cross-validation procedure was implemented into the training datasets, allowing the train dataset to validate itself for a predetermined number of folds. As a result, validation accuracies were lower than anticipated.

Apart from the overall accuracy of the models, it is also important to identify their prediction and misprediction rates. Therefore, interpretations can be made on models' capability of distinguishing classes in medicines and minerals datasets. Figures 8.9 to 8.32 show the confusion matrices of the best three models, with different validation and test accuracies of the raw and preprocessed versions for both of the train-test splits of medicines dataset, while figures 8.33 to 8.56 show same type of matrices for minerals datasets.

Figures 8.9 to 8.11 show the best 3 models' validation accuracies of the raw dataset with 70%-30% train test split.

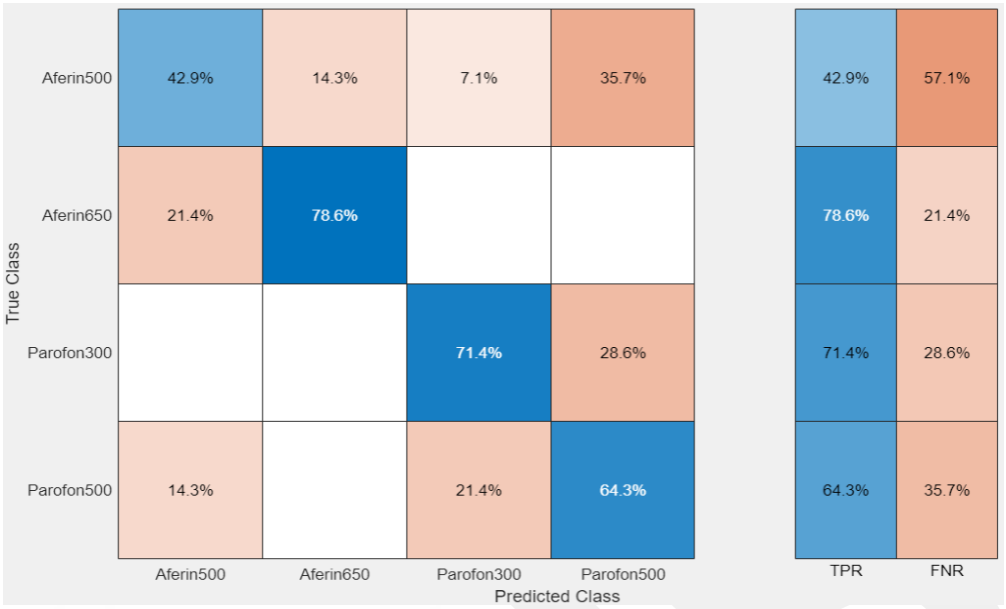

Figure 8.9 Cubic SVM (64.3% Validation Accuracy, Medicines Raw Dataset, 70%-30% train test split)

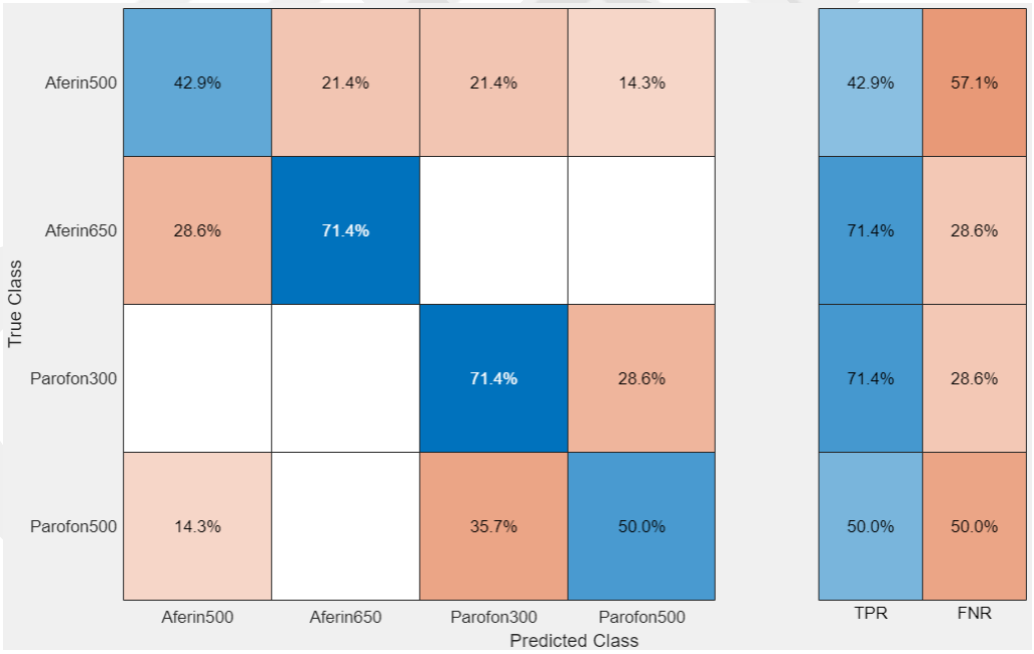

Figure 8.10 Fine KNN (58.9% Validation Accuracy, Medicines Raw Dataset, 70%-30% train test split)

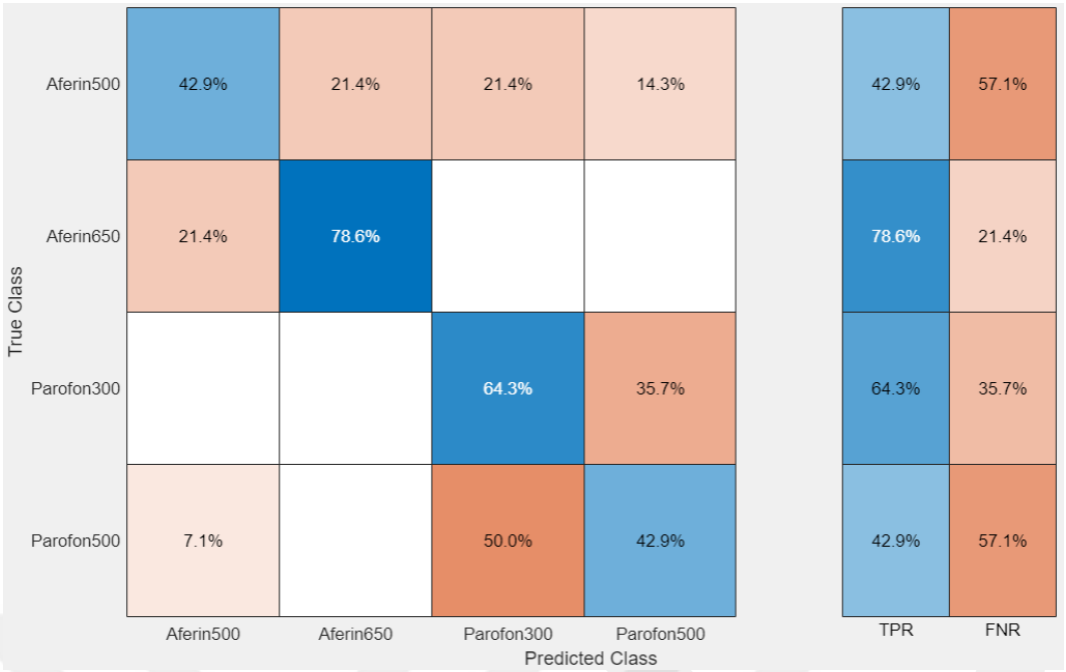

Figure 8.11 Quadratic SVM (57.1% Validation Accuracy, Medicines Raw Dataset, 70%-30% train test split)

During the validation procedure, it was shown that the top three models were more capable of differentiating between Aferin 650 mg and Parafon 300 mg than they did between Aferin 500 mg and Parafon 500 mg. It may be inferred that Parafon 300 mg and Aferin 650 mg each have unique characteristics that set them apart from one another and their various concentrations.

Figures 8.12 to 8.14 show the best 3 models' test accuracies of the raw dataset with the 70%-30% train test split.

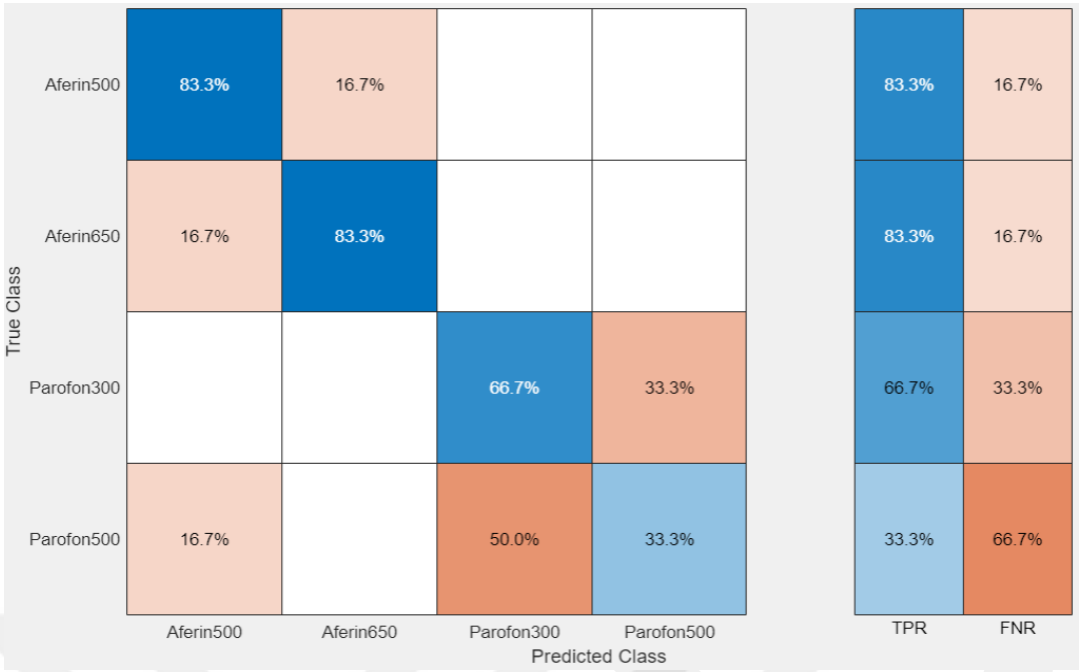

Figure 8.12 Ensemble Bagged Trees (66.7% Test Accuracy, Medicines Raw Dataset, 70%-30% train test split)

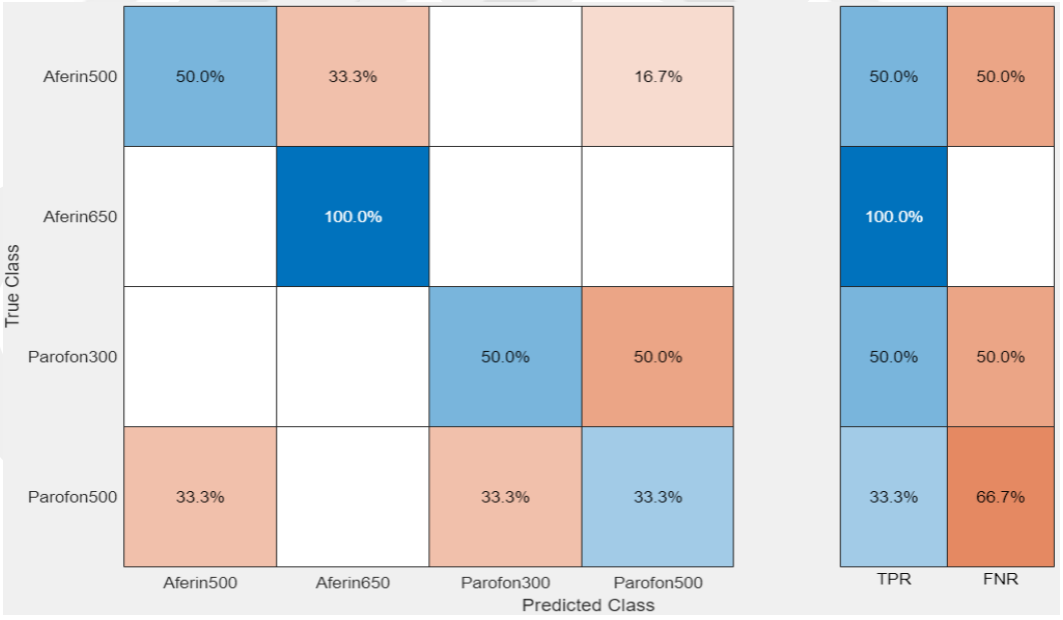

Figure 8.13 Medium Neural Network (58.3% Test Accuracy, Medicines Raw Dataset, 70%-30% train test split)

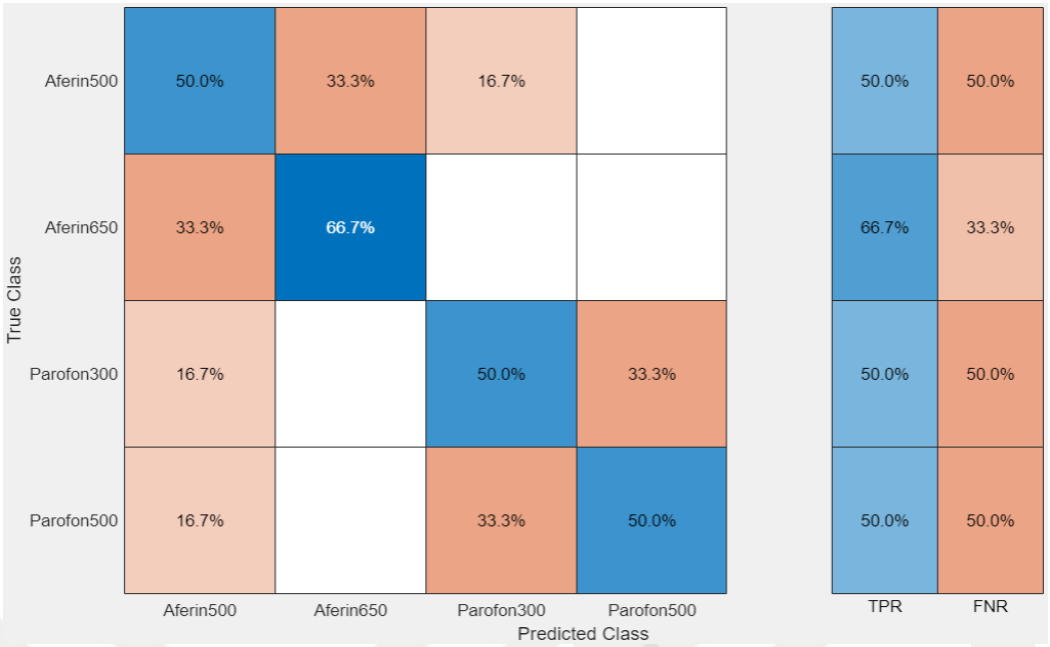

Figure 8.14 Cubic SVM (54.2% Test Accuracy, Raw Dataset, 70%-30% train test split)

Following validation, the test dataset, that was entirely new to the models, was presented. Figures 8.12 to 8.14 illustrate how poorly they classified, as it is evident. It demonstrates the need of data preprocessing for machine learning applications. Cross-validation enabled a modest improvement to be seen throughout the validation process, but it was still outside of acceptable ranges. To make a better comparison, preprocessed dataset's validation and test accuracies with the same split are shown in Figures 8.15 to 8.17 and Figures 8.18 to 8.20, respectively.

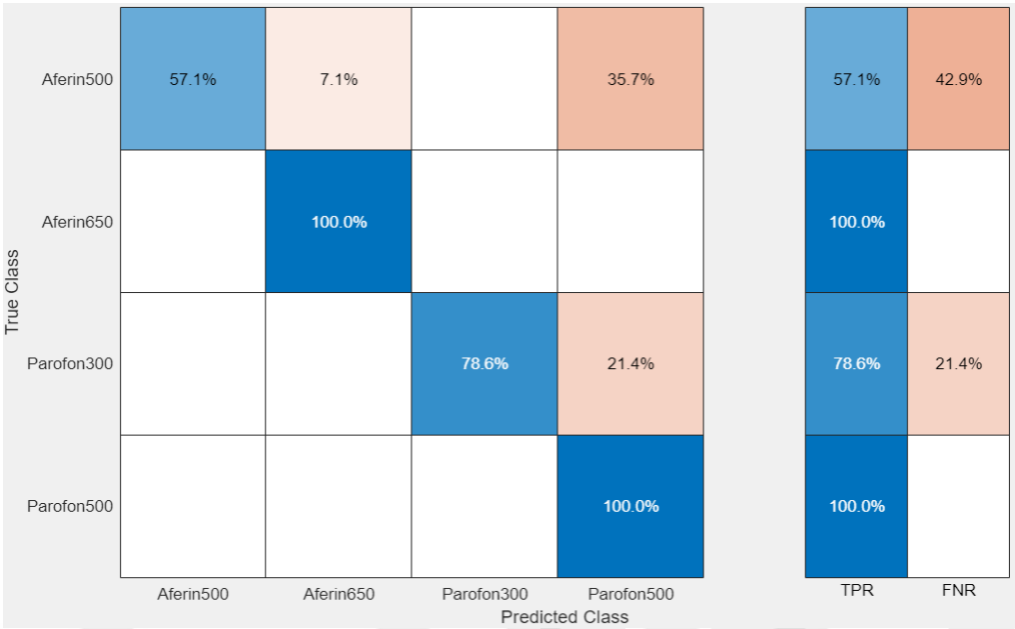

Figure 8.15 Ensemble Subspace Discriminant (83.9% Validation Accuracy, Medicines Preprocessed Dataset, 70%-30% train test split)

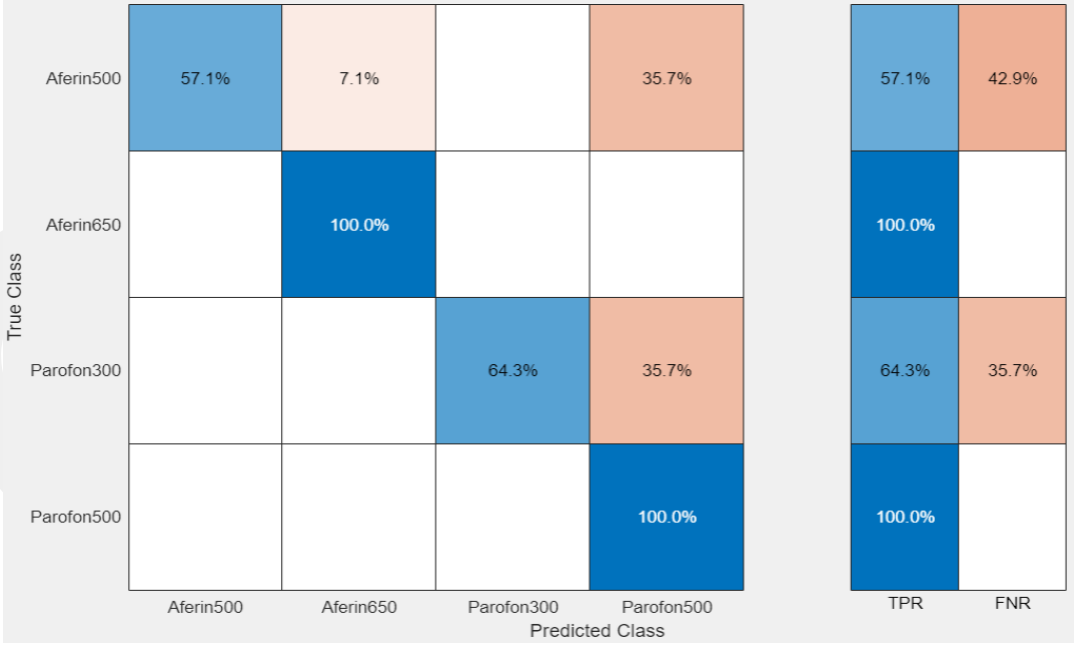

Figure 8.16 Linear Discriminant (80.4% Validation Accuracy, Medicines Preprocessed Dataset, 70%-30% train test split)

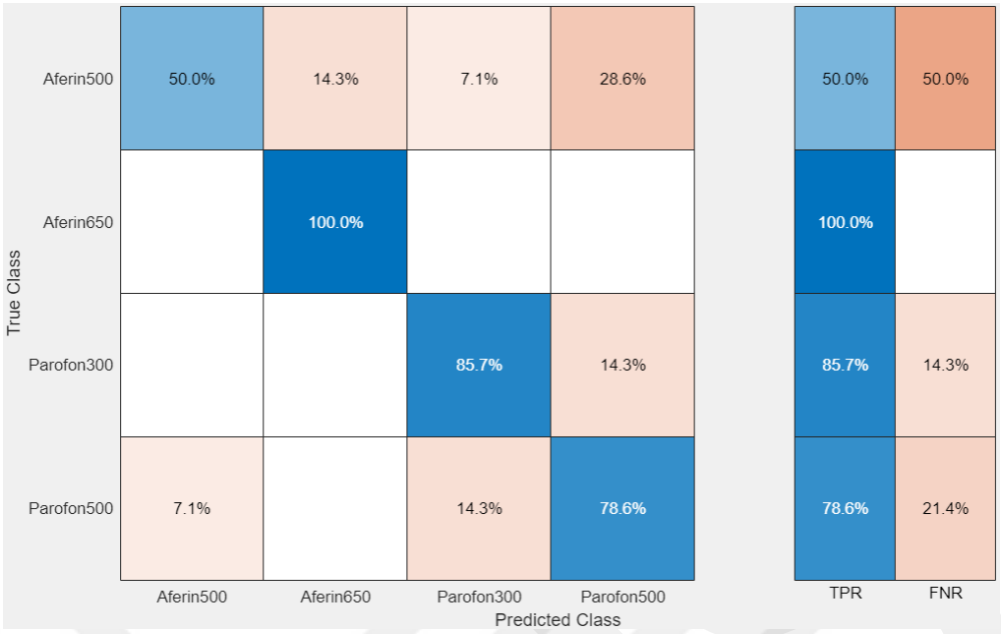

Figure 8.17 Wide Neural Network (78.6% Validation Accuracy, Medicines Preprocessed Dataset, 70%-30% train test split)

It was noted that the preprocessing significantly enhanced the validation procedure. The three top models were more successful of picking up on the distinguishing traits of Aferin and Parafon as well as their various concentrations. Although models still performed only moderately when separating Aferin 500 mg and Parafon 300 mg from their other concentrations, preprocessing increased overall accuracy.

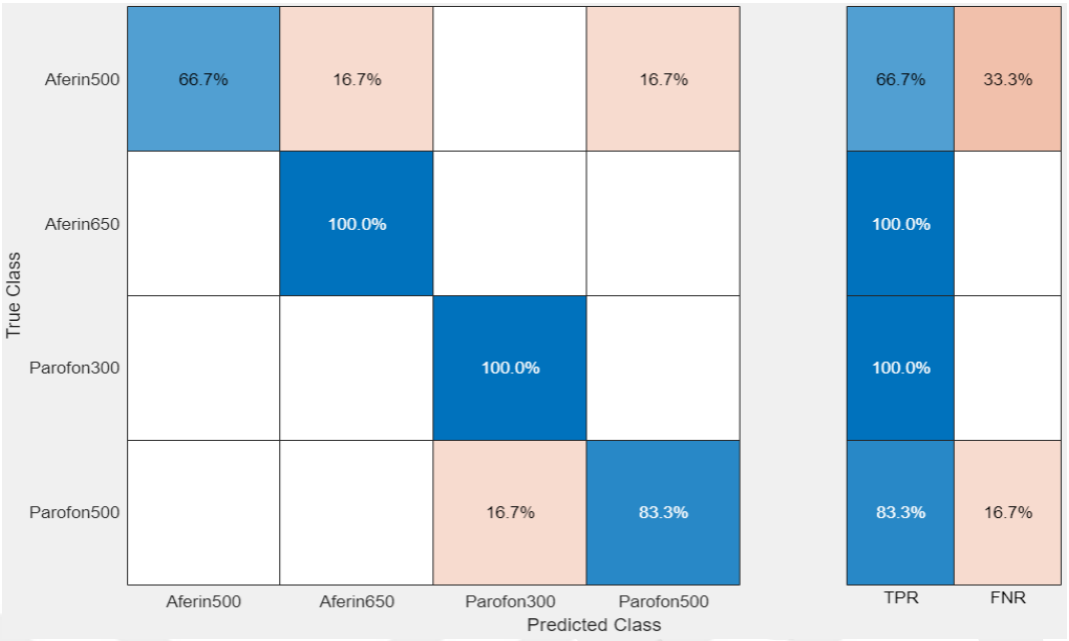

Figure 8.18 Wide Neural Network (87.5% Test Accuracy, Medicines Preprocessed Dataset, 70%-30% train test split)

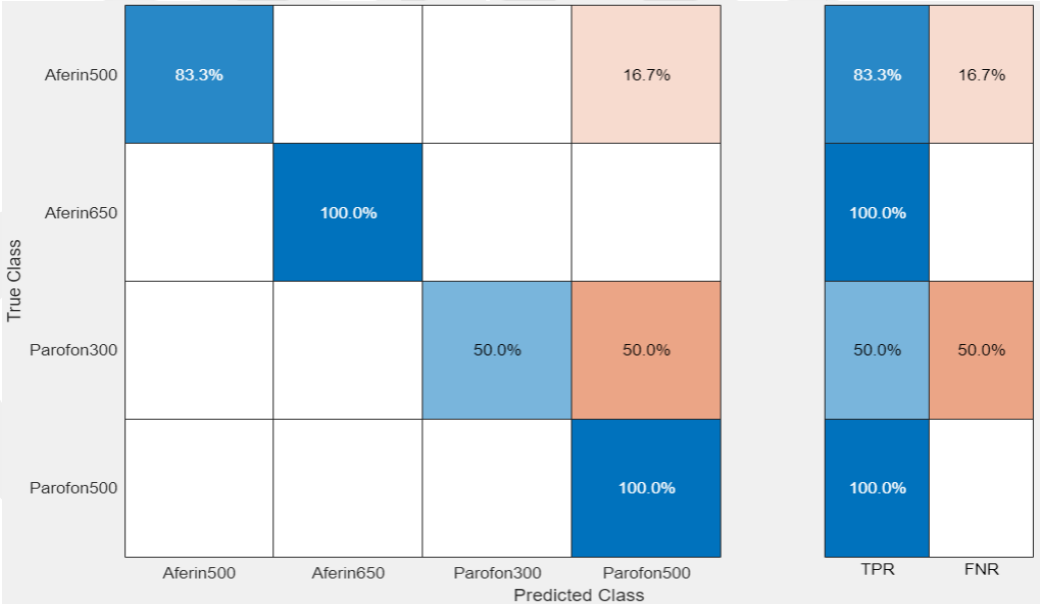

Figure 8.19 Ensemble Subspace Discriminant (83.3% Test Accuracy, Medicines Preprocessed Dataset, 70%-30% train test split)

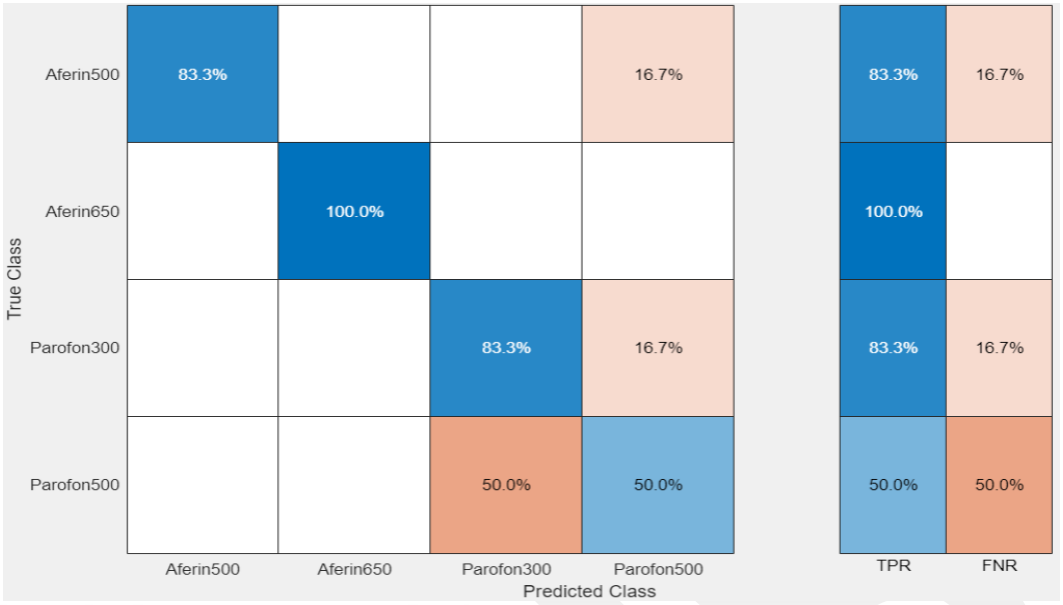

Figure 8.20 Ensemble Subspace KNN (79.2% Test Accuracy, Medicines Preprocessed Dataset, 70%-30% train test split)

Figures 8.18 to 8.20 show how the preprocessing enhanced the models' test results, resulting in greater performance in identifying Aferin and Parafon medications and their various concentrations. Preprocessing gave models to categorize Aferin and Parafon more accurately overall. In other words, the models made less mistakes while misclassifying Aferin as Parafon and vice versa. The issue of incorrectly classifying the various concentrations of Aferin and Parafon persists, nevertheless. Despite a reduction in the misclassification error at various concentrations, such as misclassifying Aferin 500 mg as Aferin 650 mg, it may still be problematic for bigger datasets in field studies. The same dataset was also exposed to same processes but with 80%-20% split. It was experimented in order to identify if different train test split would affect the machine learning classification results for both raw and preprocessed versions of the dataset.

Figures 8.21 to 8.23 show the best 3 models' validation accuracies of the raw dataset with 80%-20% train test split.

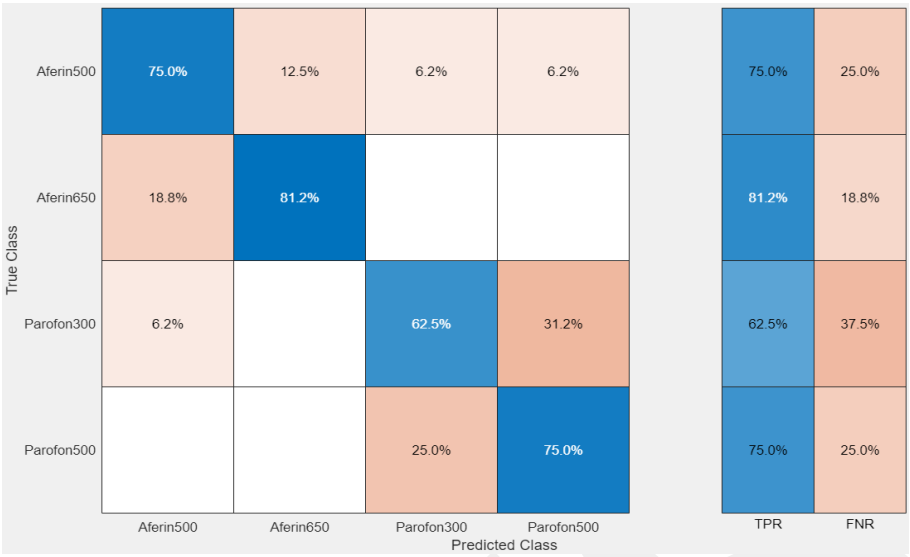

Figure 8.21 Ensemble Bagged Trees (73.4% Validation Accuracy, Medicines Raw Dataset, 80%-20% train test split)

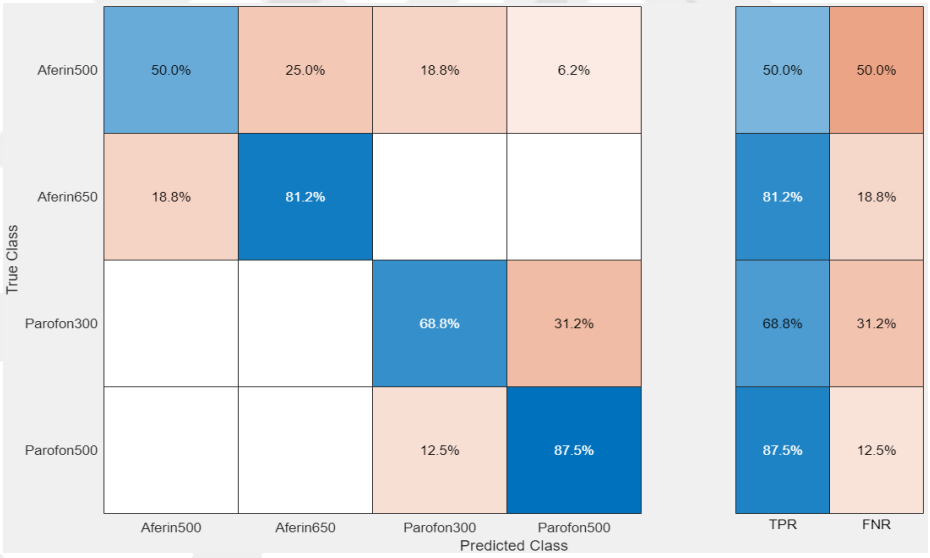

Figure 8.22 Quadratic SVM (71.9% Validation Accuracy, Medicines Raw Dataset, 80%-20% train test split)

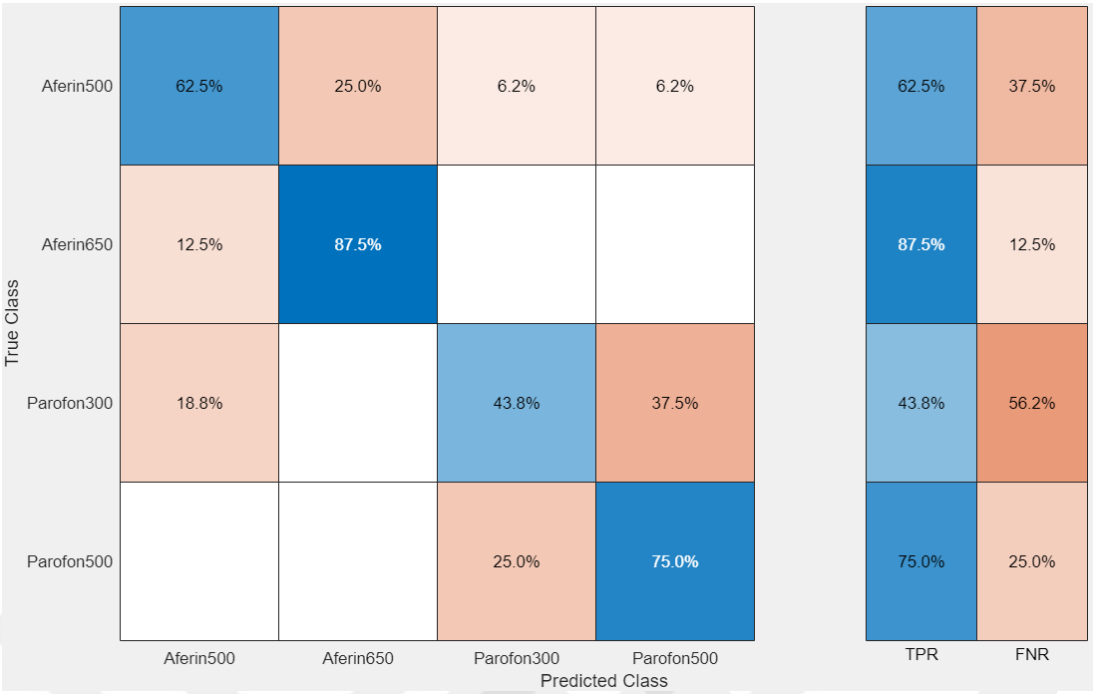

Figure 8.23 Kernel Naïve Bayes (67.2% Validation Accuracy, Medicines Raw Dataset, 80%-20% train test split)

The top three models performed better throughout the validation phase at categorizing Aferin 650 mg and Parafon 300 mg than they did at classifying Aferin 500 mg and Parafon 500 mg, similar to the 70%-30% train test split condition. It supports the hypothesis that Aferin 650 mg and Parafon 300 mg have distinctive qualities that set them apart from one another and their other concentrations. The increased number of observations for validation also improved overall validation accuracy, which was to be expected.

Figures 8.24 to 8.26 show the best 3 models' test accuracies of the raw dataset with the 80%-20% train test split.

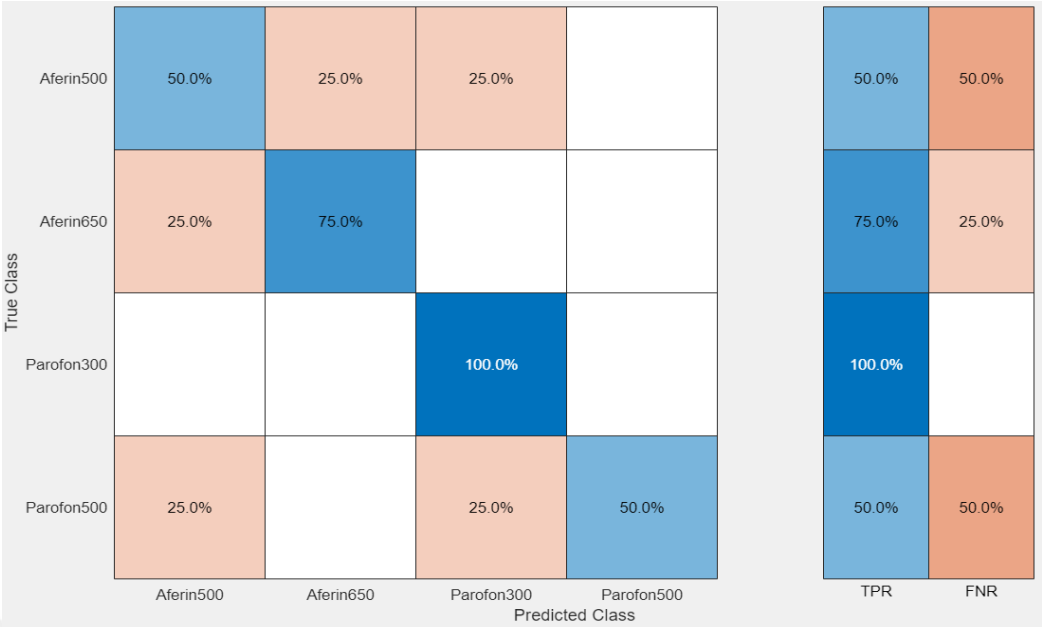

Figure 8.24 Quadratic SVM (68.8% Test Accuracy, Medicines Raw Dataset, 80%-20% train test split)

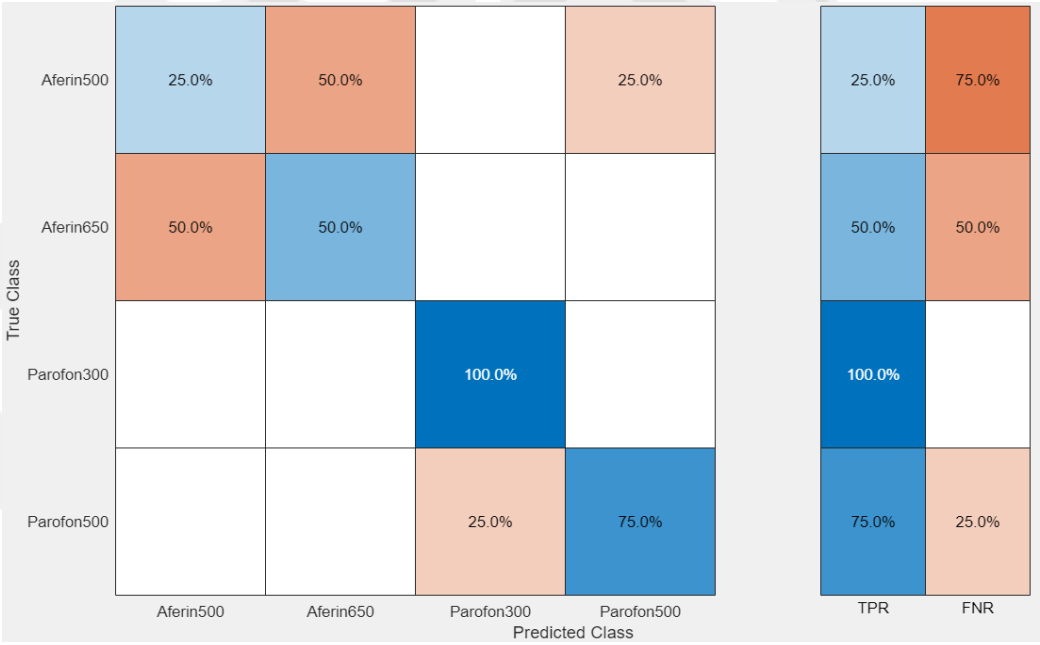

Figure 8.25 Cubic SVM (62.5% Test Accuracy, Medicines Raw Dataset, 80%-20% train test split)

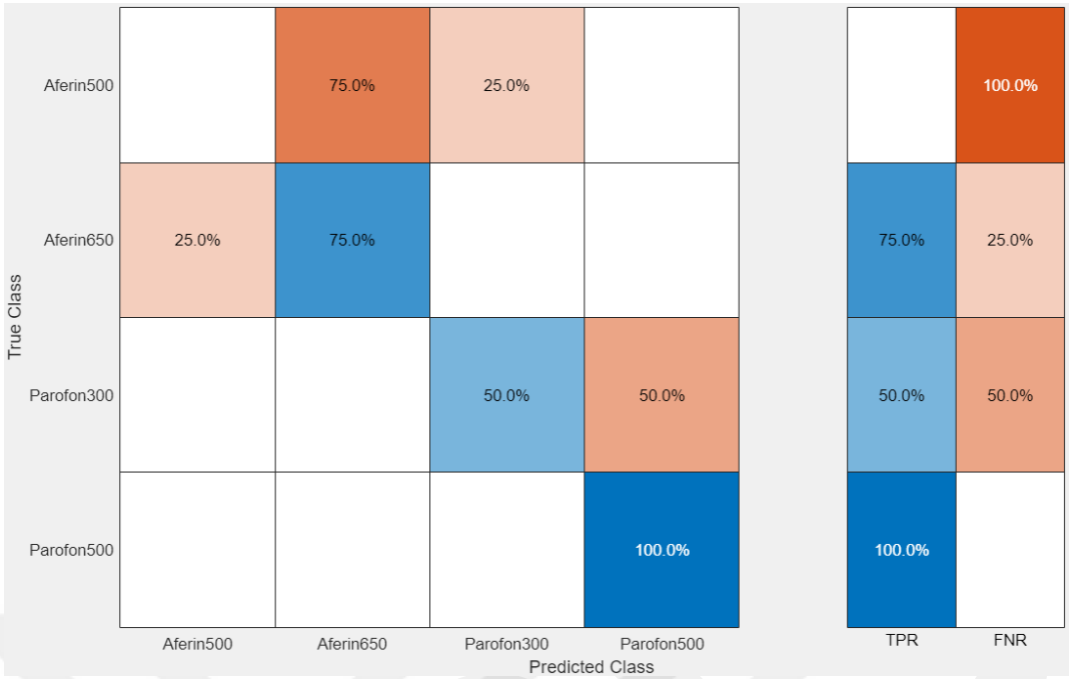

Figure 8.26 Wide Neural Network (56.2% Test Accuracy, Medicines Raw Dataset, 80%-20% train test split)

It was shown that the accuracy of raw data with a 70%–30% split did not significantly improve when the amount of the validation data was increased from 70% to 80%. However, with a split of 70%–30%, both overall accuracy and true positive rates for each class are still only mediocre and underperforming when compared to the corresponding values of preprocessed data. To make a better comparison, preprocessed dataset's validation and test accuracies with 80%-20% are shown in Figures 8.27 to 8.29 and Figures 8.30 to 8.32, respectively.

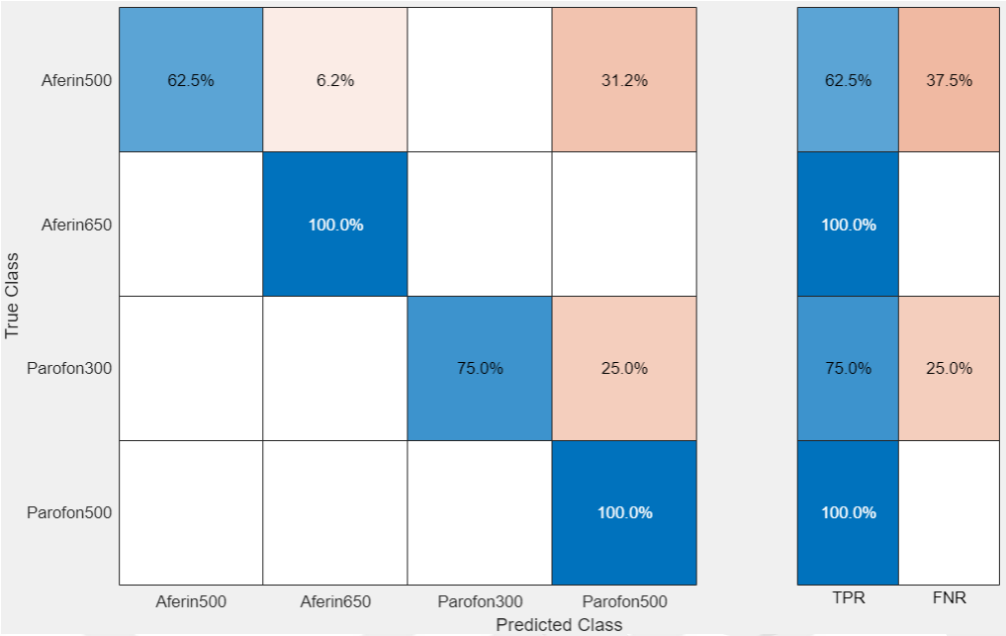

Figure 8.27 Linear Discriminant (84.4% Validation Accuracy, Medicines Preprocessed Dataset, 80%-20% train test split)

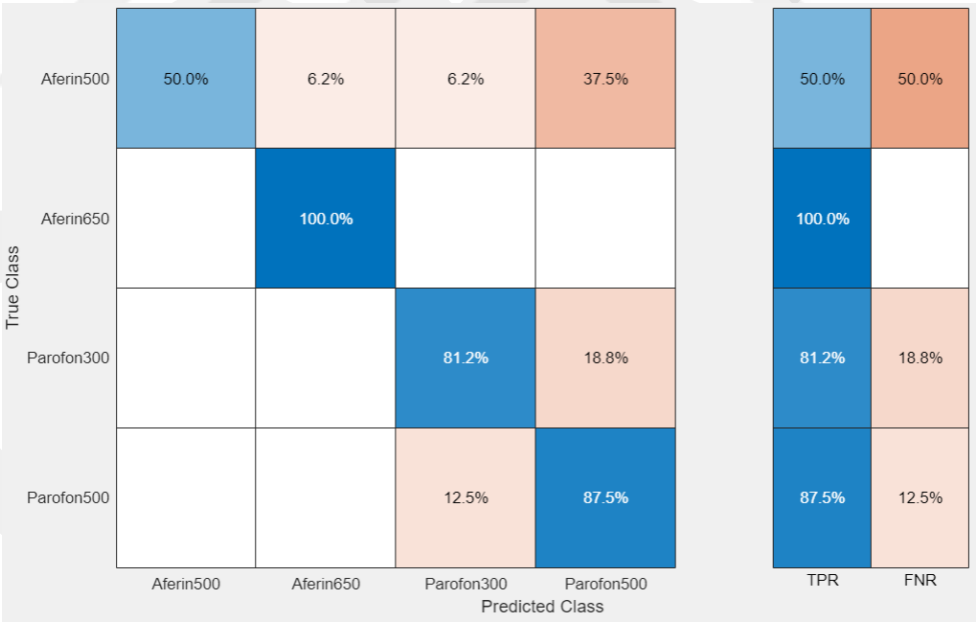

Figure 8.28 Wide Neural Network (79.7% Validation Accuracy, Medicines Preprocessed Dataset, 80%-20% train test split)

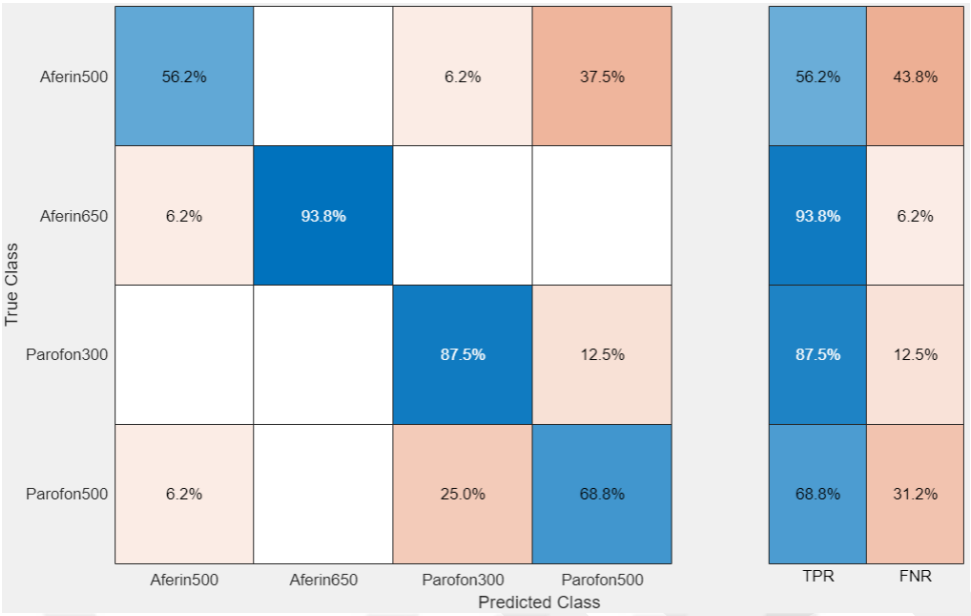

Figure 8.29 Cosine KNN (76.6% Validation Accuracy, Medicines Preprocessed Dataset, 80%-20% train test split)

Figures 8.27 to 8.29 show that overall validation accuracies of the top three models were increased as a result of preprocessing as compared to raw dataset. Aferin 500 mg and Parafon 300 mg continue to be misclassified, which is comparable to the preprocessed dataset's validation accuracy for a 70%–30% split version. It should be highlighted that the Cosine KNN model, the third-best model, performed better when it came to classifying Parafon 300mg. However, it performs worse when categorizing the other 3 classifications. Figures 8.30 to 8.32 show the best 3 models' test accuracies of the preprocessed dataset with the 80%-20% train test split.

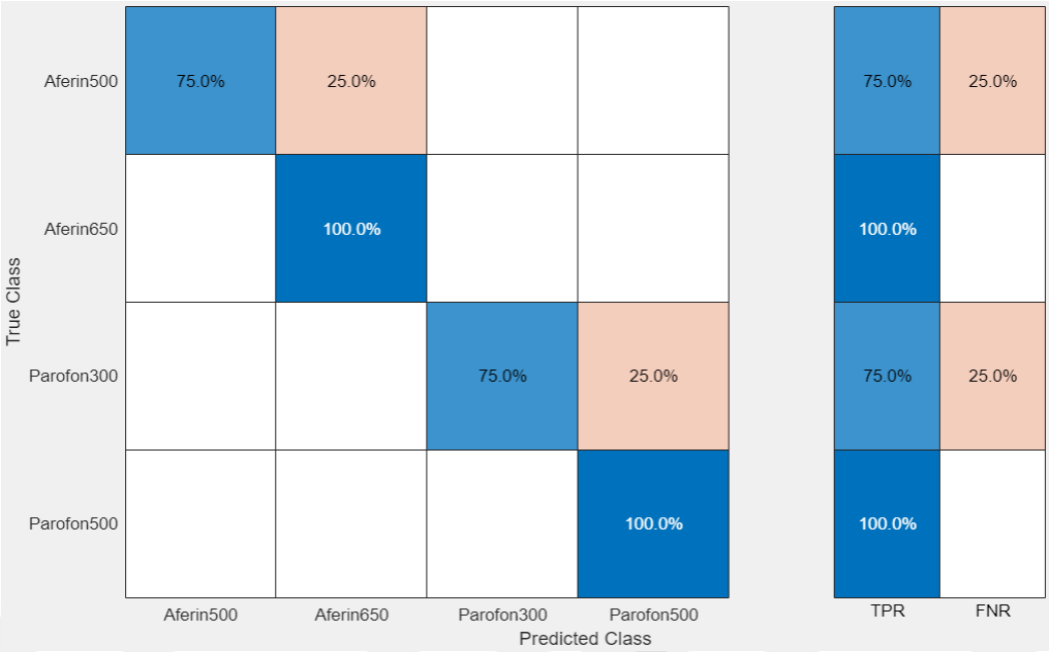

Figure 8.30 Linear Discriminant (87.5% Test Accuracy, Medicines Preprocessed Dataset, 80%-20% train test split)

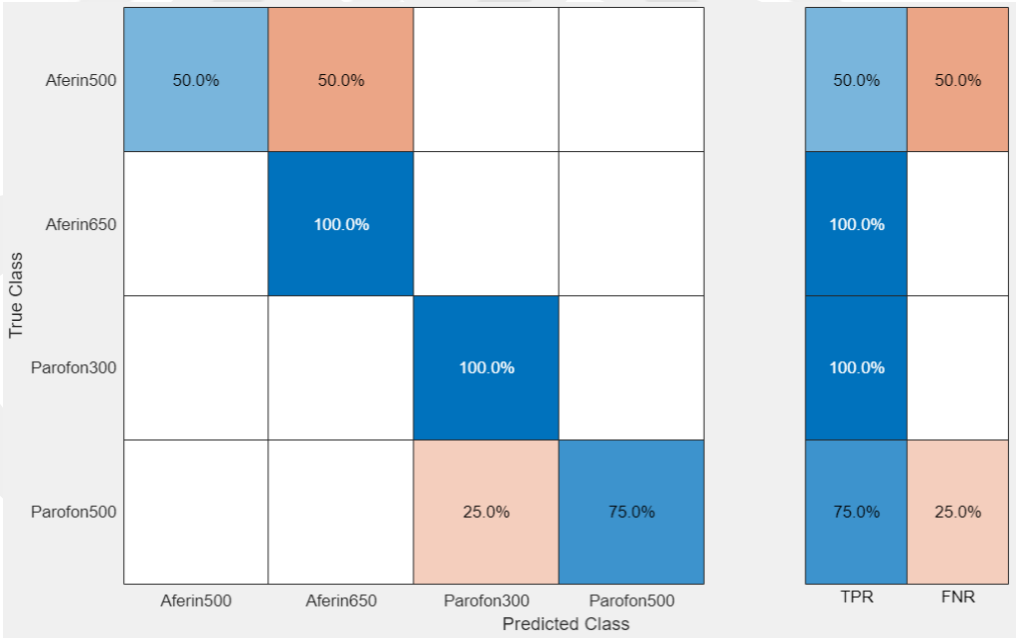

Figure 8.31 Ensemble Subspace KNN (81.2% Test Accuracy, Medicines Preprocessed Dataset, 80%-20% train test split)

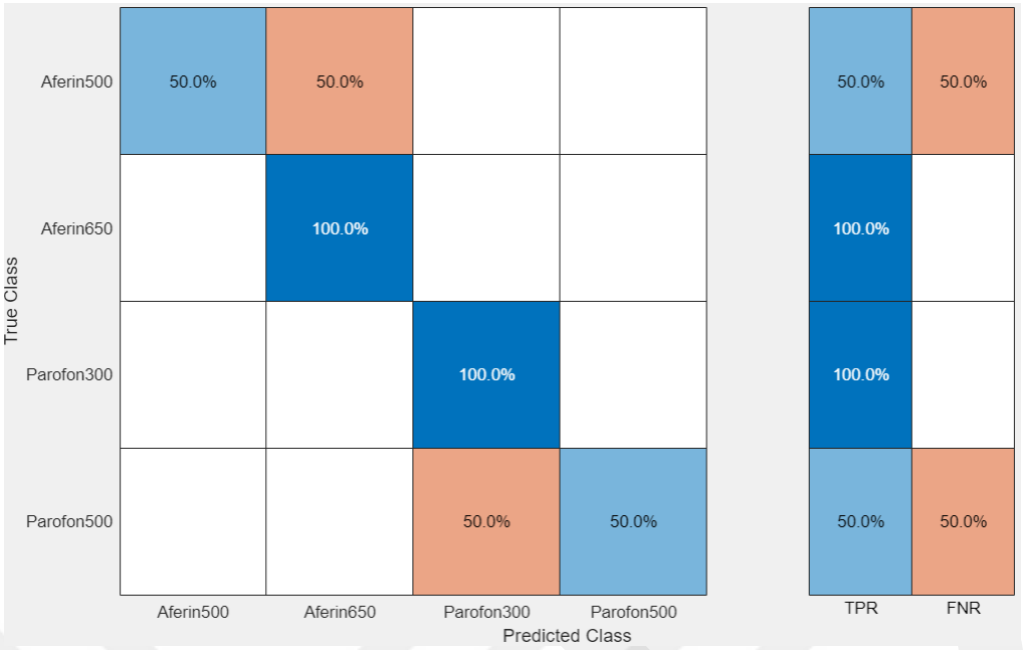

Figure 8.32 Cosine KNN (75% Test Accuracy, Medicines Preprocessed Dataset, 80%-20% train test split)

Aferin 650 mg was correctly categorized by all three models. In contrast, the first model, Linear Discriminant model, categorized Parafon 500 mg, while the second and third models, Ensemble Subspace KNN model and Cosine KNN model, correctly identified Parafon 300 mg, without any error.

From this point on, results of the minerals dataset are investigated. Figures 8.33 to 8.35 show the best 3 models' validation accuracies of the raw dataset with 70%-30% train test split, for minerals dataset.

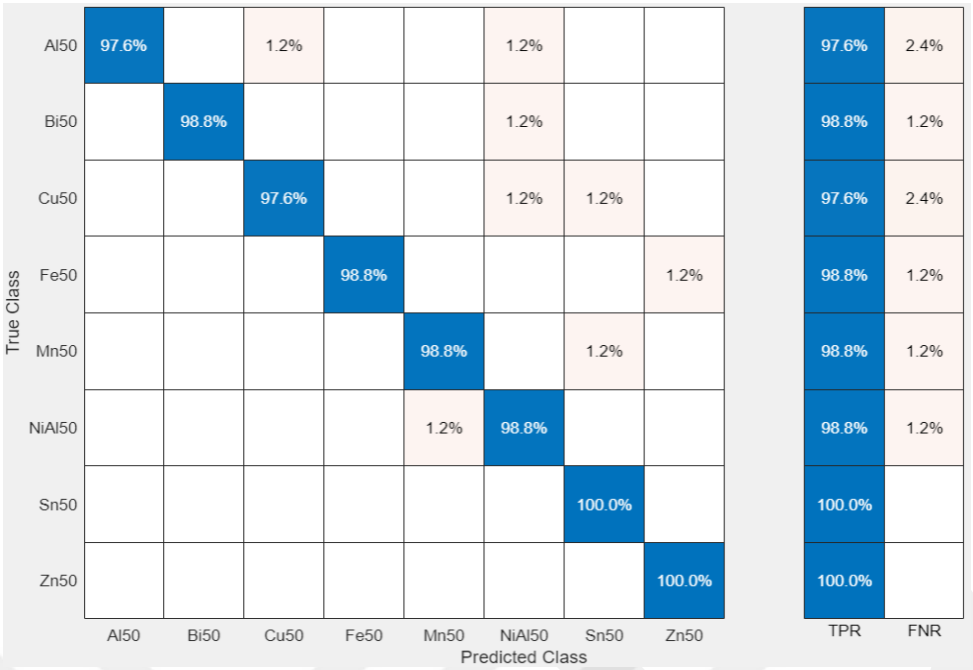

Figure 8.33 Quadratic SVM (98.8% Validation Accuracy, Minerals Raw Dataset, 70%-30% train test split)

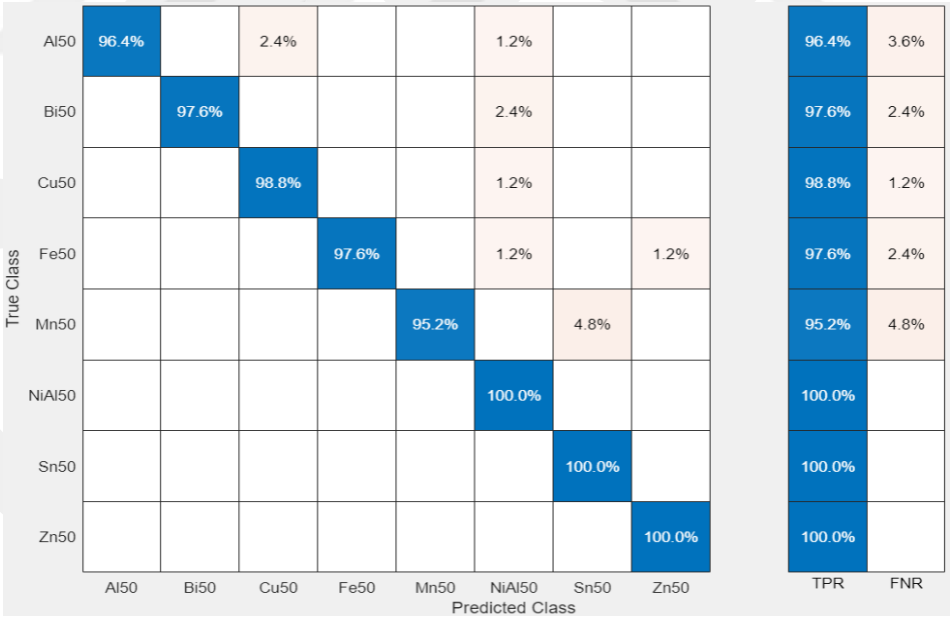

Figure 8.34 Linear SVM (98.2% Validation Accuracy, Minerals Raw Dataset, 70%-30% train test split)

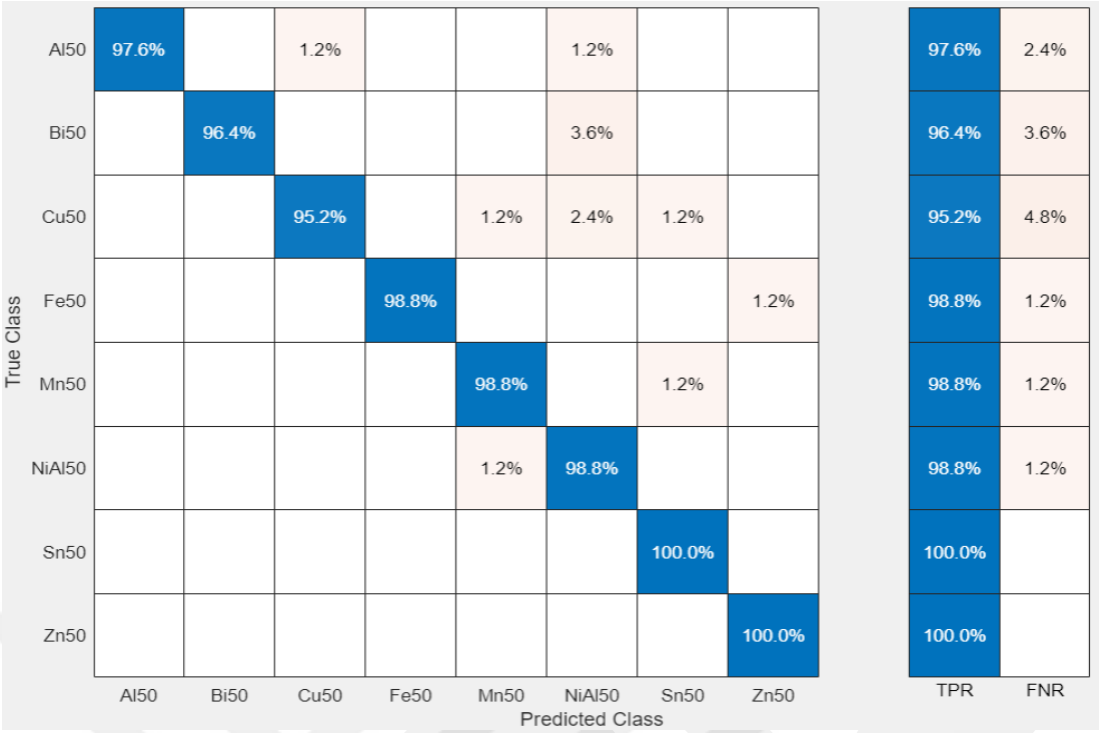

Figure 8.35 Cubic SVM (98.2% Validation Accuracy, Minerals Raw Dataset, 70%-30% train test split)

As it is evident in the Figures 8.32 to 8.34, Sn50 and Zn50 have distinguishing spectroscopy characteristics which facilitated the models to classify them without any error. Although having the high percent true positive rates, other 6 mineral have similar characteristics to prevent them being distinguished with 100% success. Also, it is seen that SVM models are highly capable at classification, despite 8 different classes.

Figures 8.36 to 8.38, show the best 3 models' test accuracies of the raw dataset with the 70%-30% train test split.

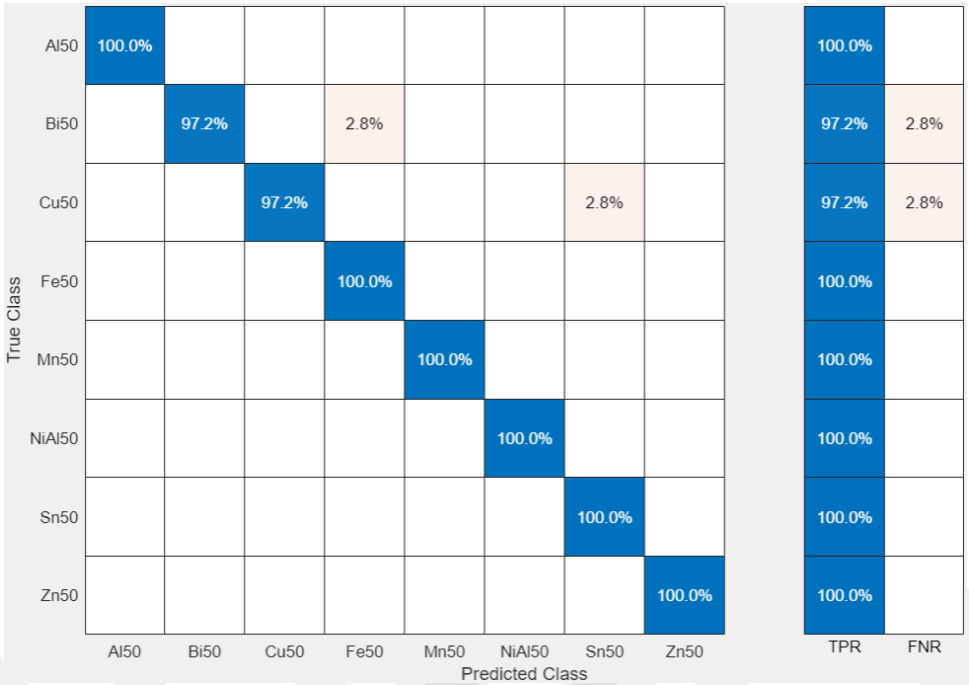

Figure 8.36 Quadratic SVM (99.3% Test Accuracy, Minerals Raw Dataset, 70%-30% train test split)

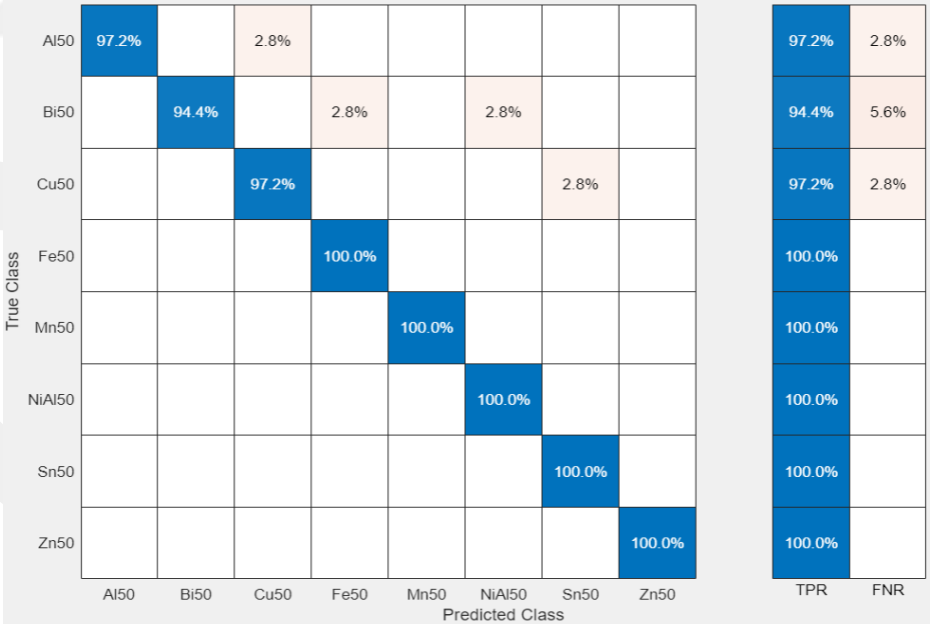

Figure 8.37 Cubic SVM (98.6% Test Accuracy, Minerals Raw Dataset, 70%-30% train test split)

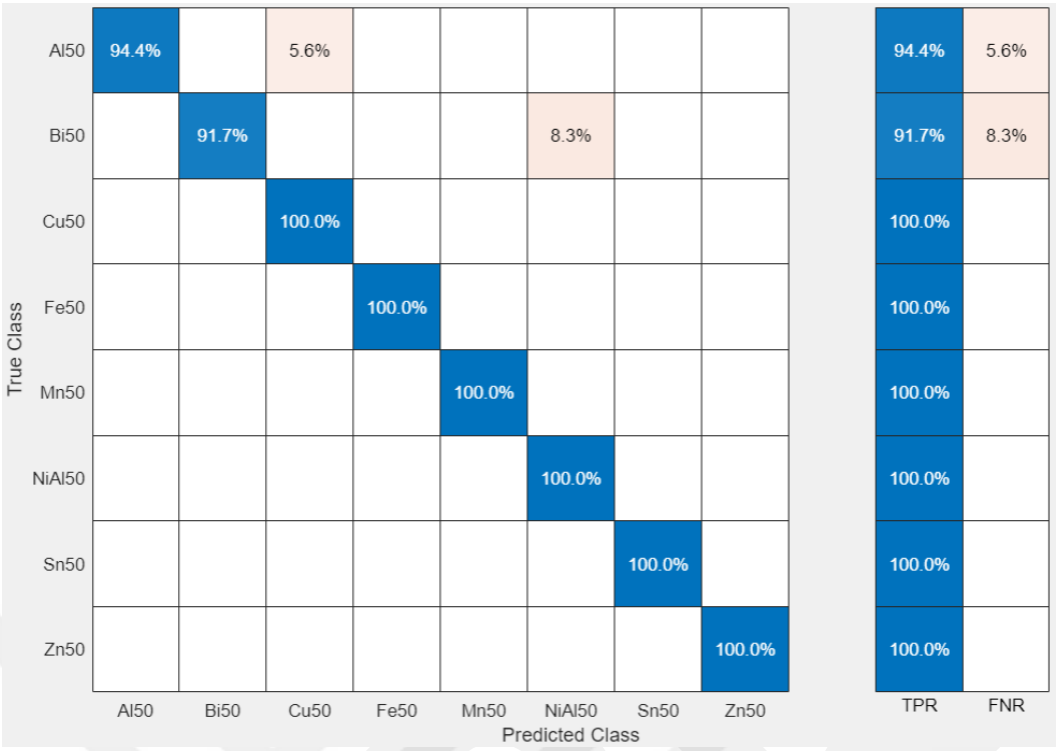

Figure 8.38 Linear SVM (98.3% Test Accuracy, Minerals Raw Dataset, 70%-30% train test split)

All three SVM models; quadratic, cubic and linear SVM models; scored much higher accuracies with test data. It must be noted that, test accuracies of the models resulted a little bit higher than their respective validation accuracies. This outcome is sourced from the cross-validation process. After standard training procedure is completed, cross-validation separates the training data into certain number of folds and it performs validation in each fold. Thus, the overall validation accuracy is obtained and it is always lower than training accuracy. When validation accuracy is compared to test accuracy, it is often found to be lower than test accuracy. Returning to the main point; the models failed to achieve only 3 minerals, Al5, Bi5, and Cu50, with 100% accuracy. However, their errors seem in acceptable limits.

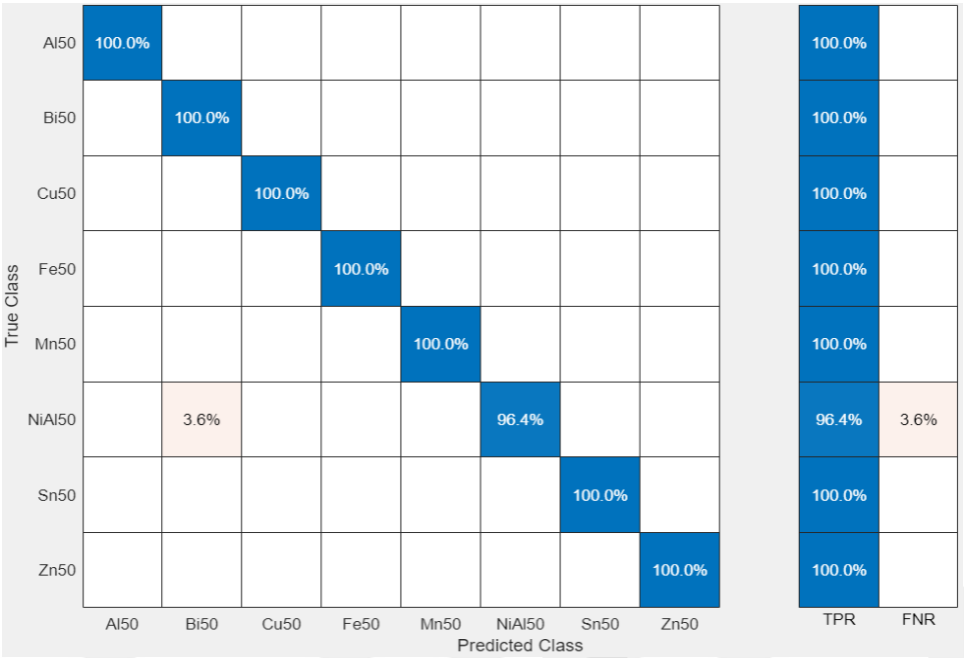

Figure 8.39 Quadratic SVM (99.6% Validation Accuracy, Minerals Preprocessed Dataset, 70%-30% train test split)

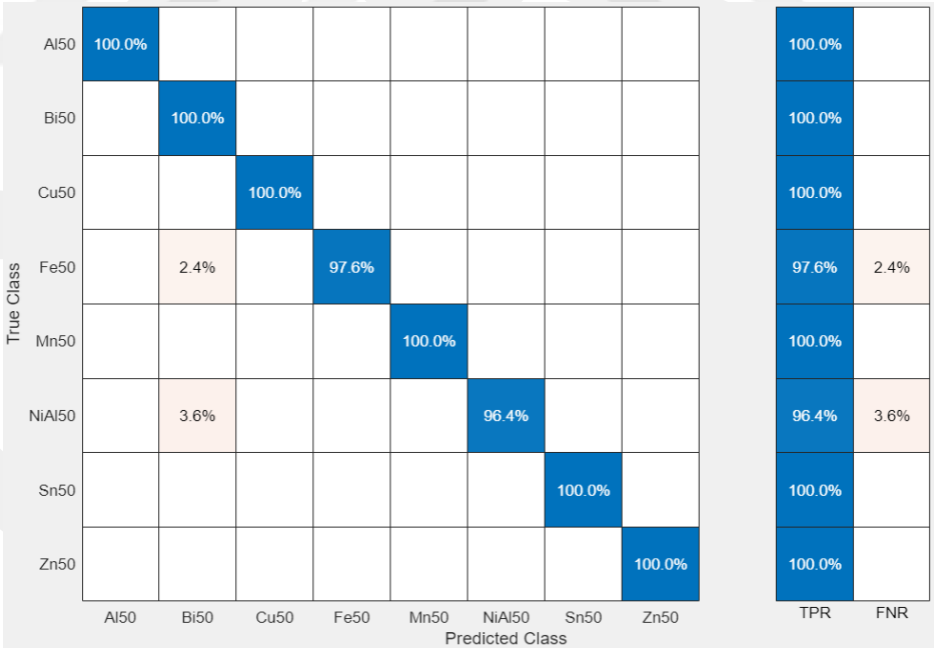

Figure 8.40 Cubic SVM (99.3% Validation Accuracy, Minerals Preprocessed Dataset, 70%-30% train test split)

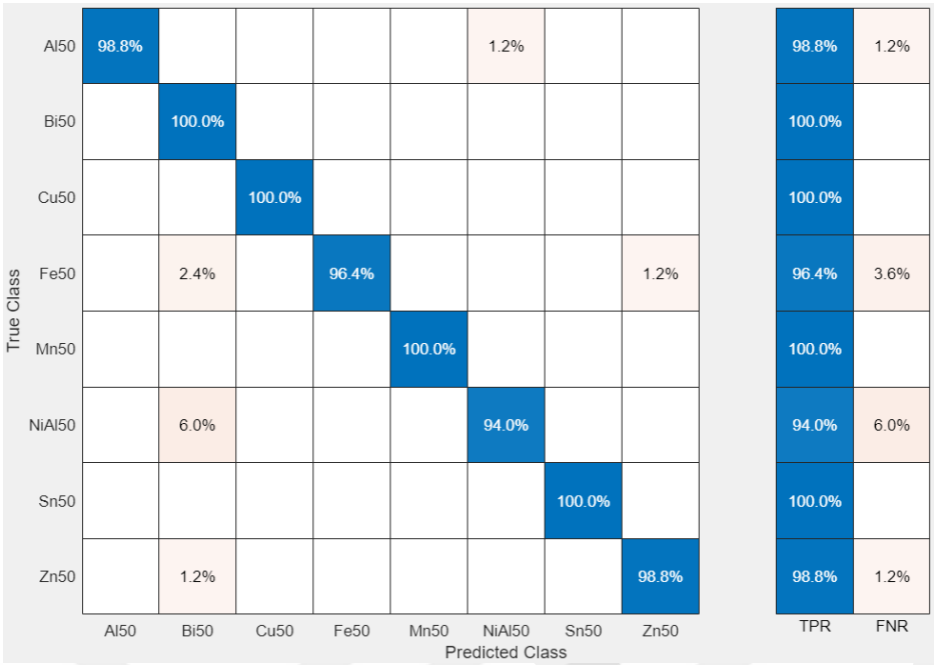

Figure 8.41 Wide Neural Network (99.0% Validation Accuracy, Minerals Preprocessed Dataset, 70%-30% train test split)

It is seen that preprocessing performed fine tuning on Wide NN, Quadratic and Cubic SVM models while causing a minor accuracy drop, to 94.9%, for Linear SVM model. Therefore, necessity of preprocessing on SVM models might remain a debatable issue. To have a clear idea on the situation, testing accuracies were investigated next.

Figures 8.42 to 8.44 show the best three models' test accuracies of preprocessed dataset with the 70%-30% train test split.

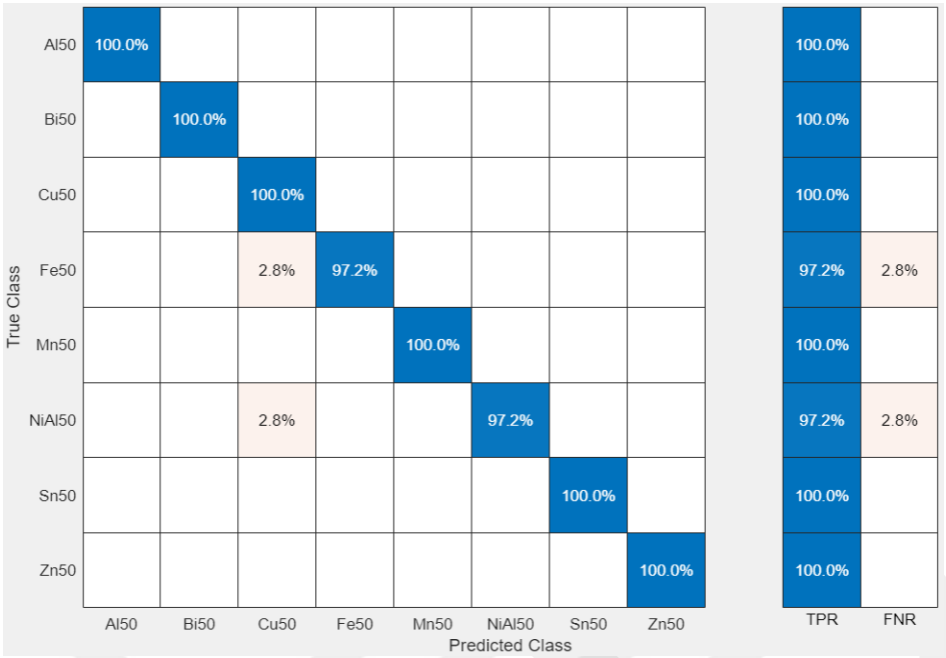

Figure 8.42 Wide NN (99.7% Test Accuracy, Minerals Preprocessed Dataset, 70%-30% train test split)

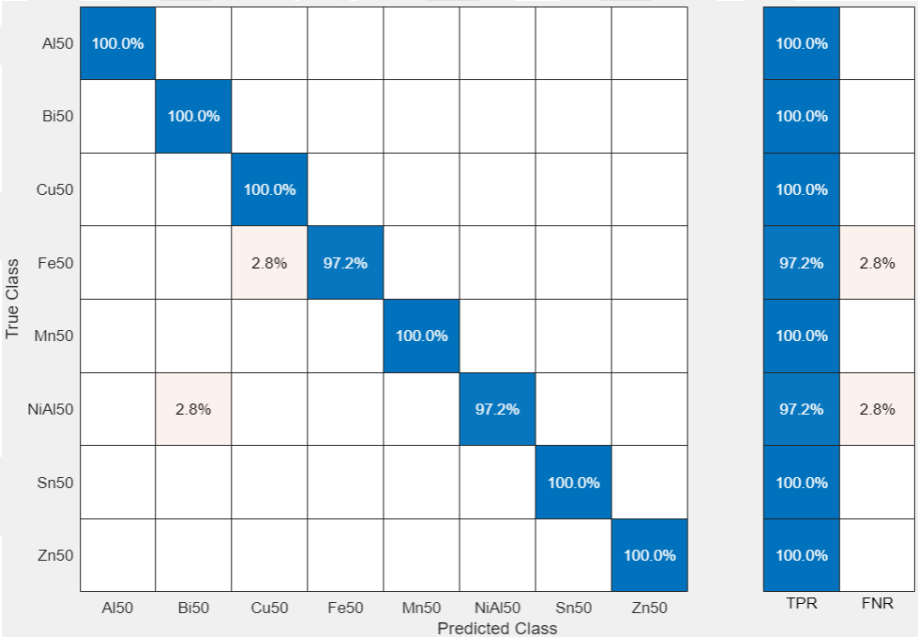

Figure 8.43 Quadratic SVM (99.3% Test Accuracy, Minerals Preprocessed Dataset, 70%-30% train test split)

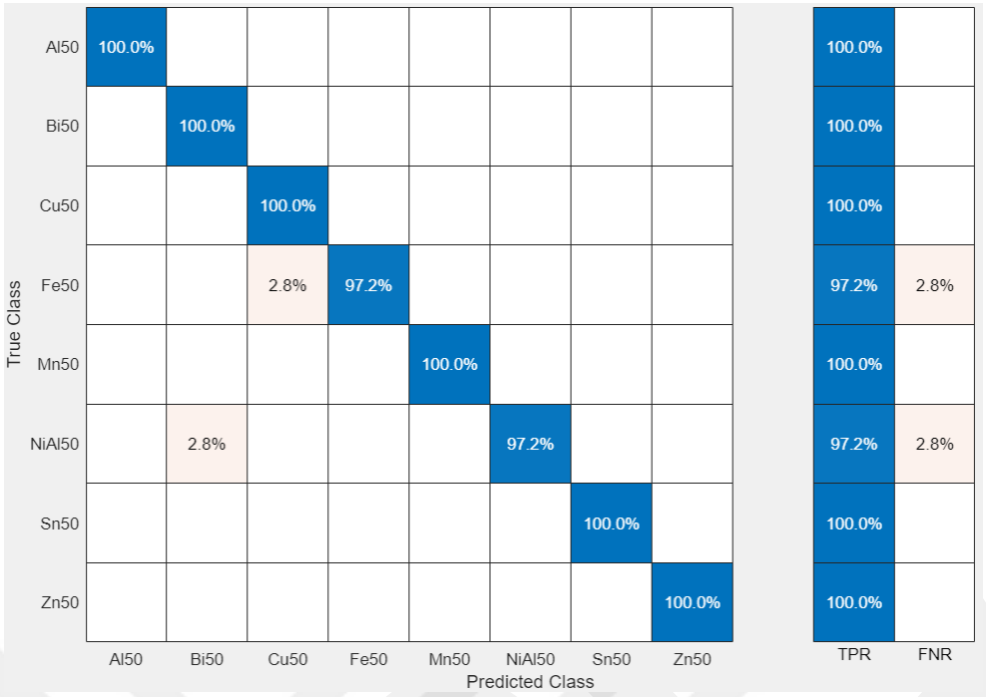

Figure 8.44 Cubic SVM (99.3% Test Accuracy, Minerals Preprocessed Dataset, 70%-30% train test split)

As it is seen in the figures above, preprocessing made a negligible impact on the models' test accuracies in comparison with both their respective both test accuracies of raw dataset and validation accuracies of preprocessed dataset. From this view, effect preprocessing becomes nearly ignorable for the models, which have already achieved highest performance. To have a better understanding on the issue, 80%-20% train-test split of the mineral dataset will be examined next. Therefore, 80%-20% train test split of minerals dataset, raw and preprocessed versions, were investigated from figures 8.44 to 8.55.

Figures 8.45 to 8.47 show the best three models' validation accuracies of raw dataset with 80%-20% split.

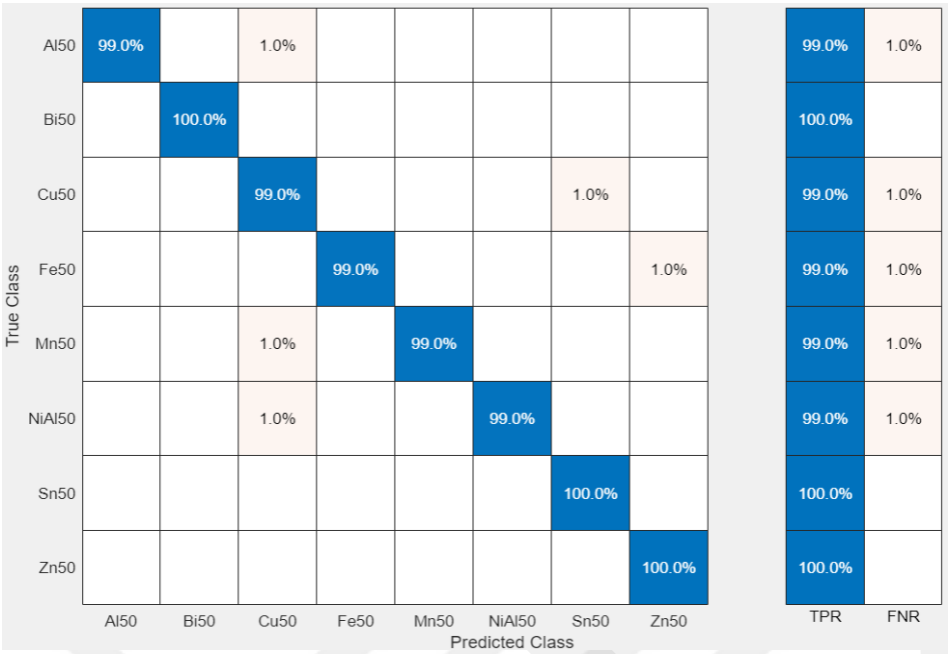

Figure 8.45 Cubic SVM (99.3% Validation Accuracy, Minerals Raw Dataset, 80%-20% train test split)

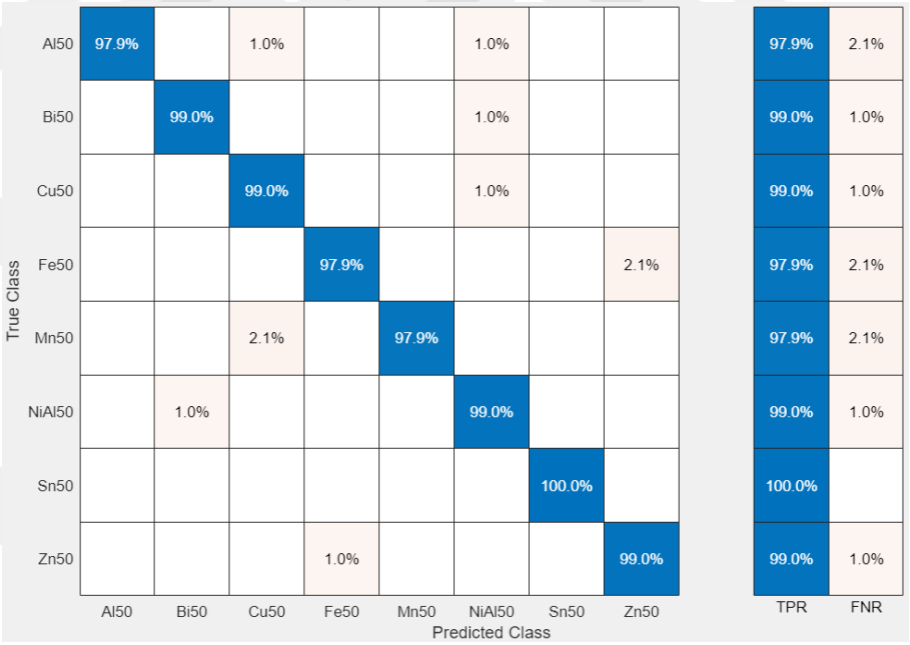

Figure 8.46 Quadratic SVM (98.7% Validation Accuracy, Minerals Raw Dataset, 80%-20% train test split)

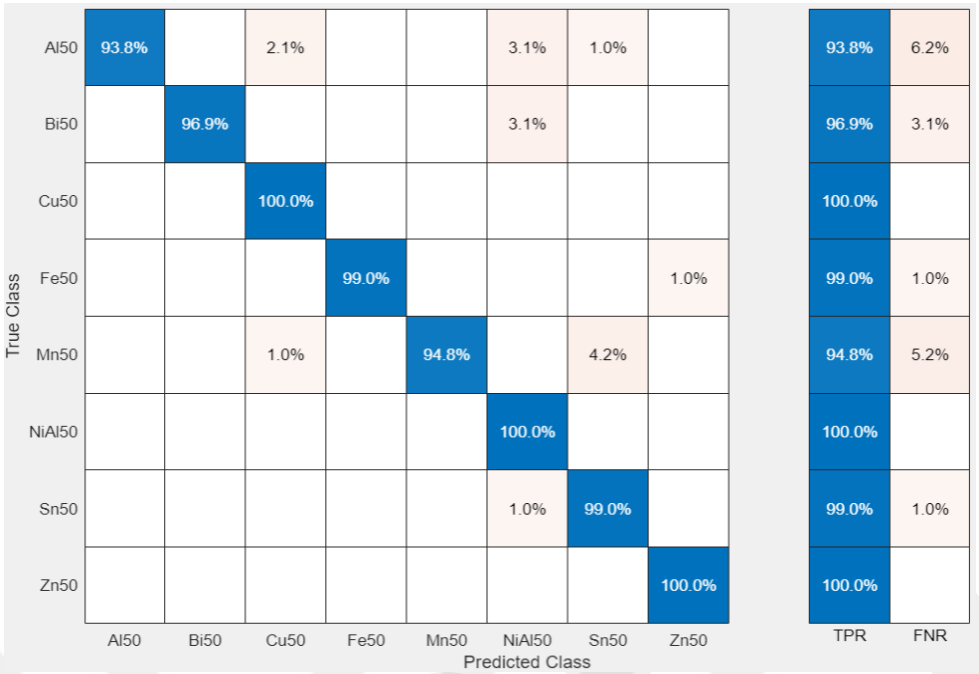

Figure 8.47 Linear SVM (98.6% Validation Accuracy, Minerals Raw Dataset, 80%-20% train test split)

As it is seen from the figures above, increase of training dataset from 70% to 80% did not make a considerable impact on validation accuracy. Therefore, it can be deduced that 70% of dataset suffices for Cubic, Quadratic and Linear SVM models to effectively classify different minerals. Thus, extra amount of 10% of data does not improve overall performance but only increases the workload for these models. However, the test accuracies were investigated to have better understanding of the issue at hand.

Figures 8.47 to 8.49 show the best three models' test accuracies of raw dataset with 80%-20% split.

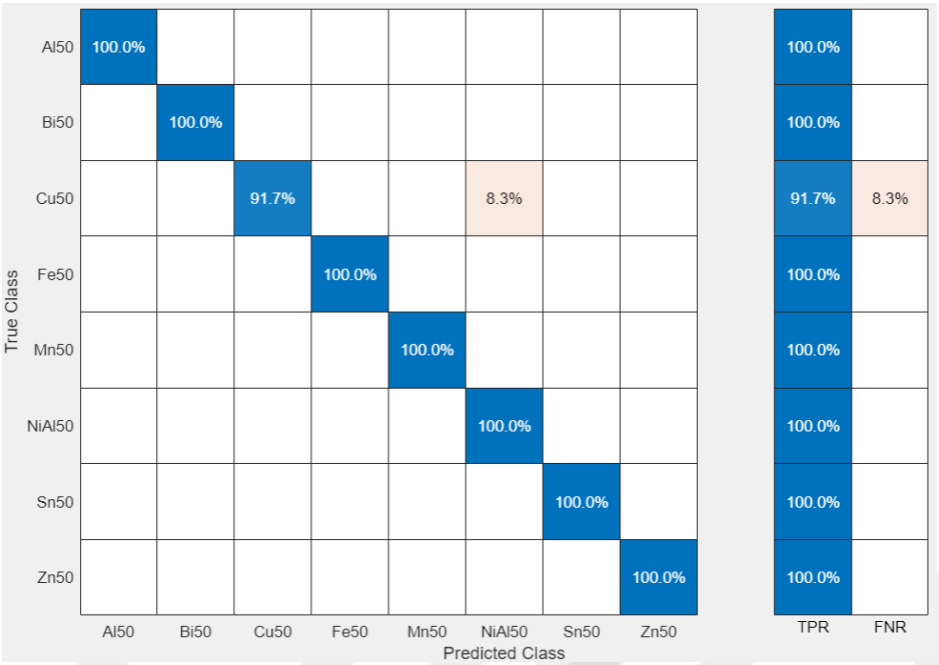

Figure 8.48 Quadratic SVM (99.0%Test Accuracy, Minerals Raw Dataset, 80%-20% train test split)

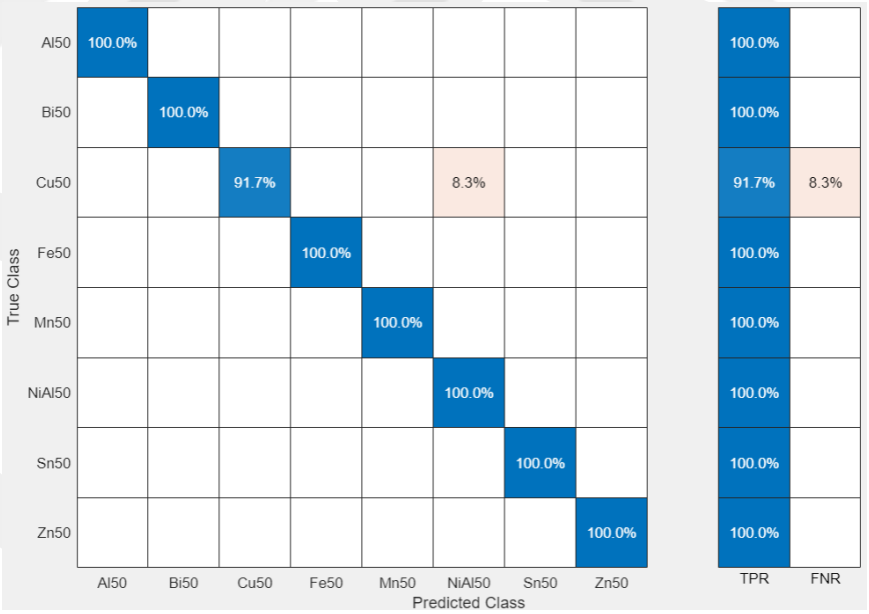

Figure 8.49 Cubic SVM (99.0% Test Accuracy, Minerals Raw Dataset, 80%-20% train test split)

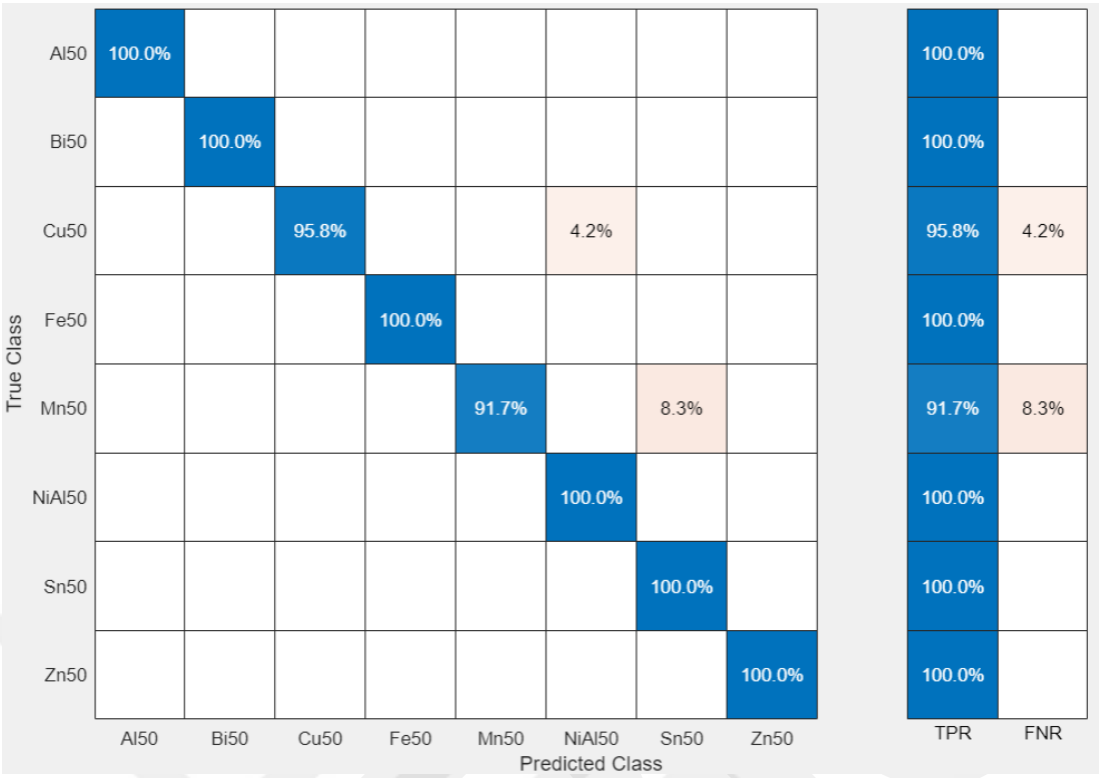

Figure 8.50 Linear SVM (98.4% Test Accuracy, Minerals Raw Dataset, 80%-20% train test split)

Similary to the validation accuracies of raw dataset with 80%-20% split, the overall test accuracies of the models remained almost same with their 70%-30% split version. Although, a chronical misclassification of Cu50 as NiAl50 is revealed. Therefore, it can be elicited that these models may have a trait which deter them to bisect Cu50 and NiAl50 with 100% success. From now on, the effects of preprocessing on validation ant test accuracies for 80%-20% train-test split of data were investigated.

Figures 8.51 to 8.53 show the best three models' validation accuracies of preprocessed dataset with 80%-20% split.

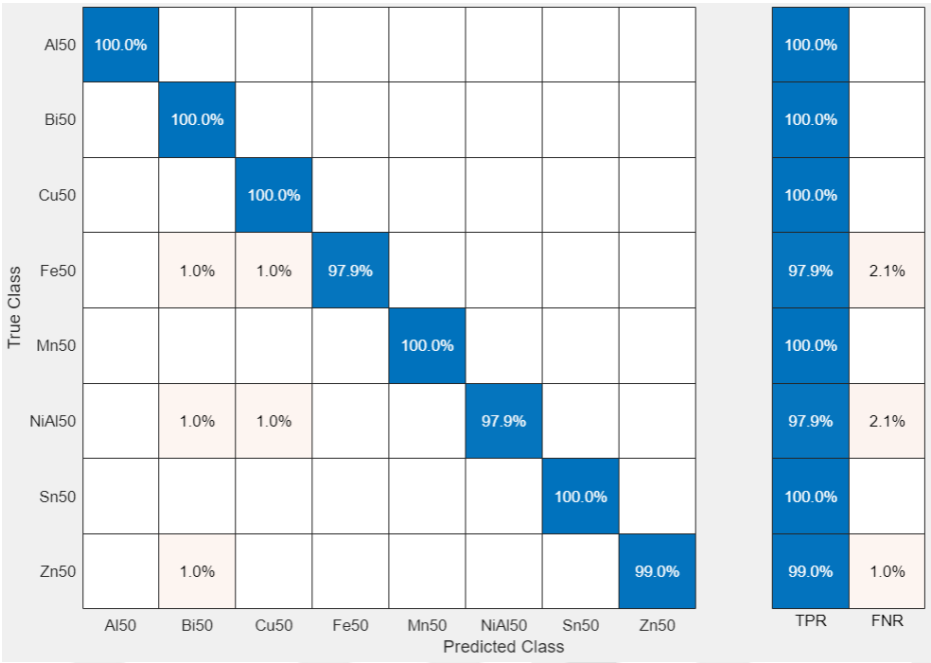

Figure 8.51 Quadratic SVM (99.3% Validation Accuracy, Minerals Preprocessed Dataset, 80%-20% train test split)

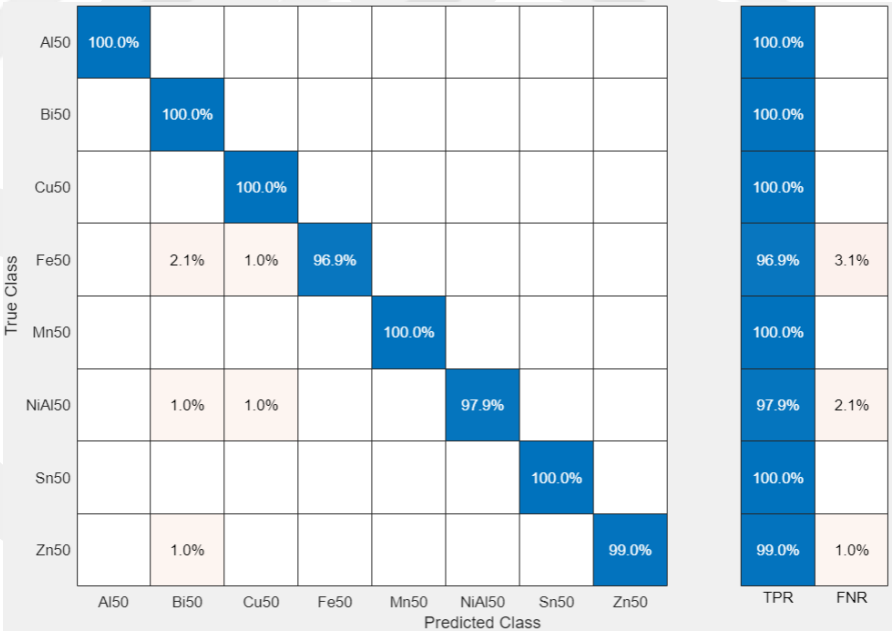

Figure 8.52 Cubic SVM (99.2% Validation Accuracy, Minerals Preprocessed Dataset, 80%-20% train test split)

Figure 8.53 Linear Discriminant (98.7% Validation Accuracy, Minerals Preprocessed Dataset, 80%-20% train test split)

From the above figures 8.51 to 8.53, it was observed that Quadratic and Cubic SVM models did not show a significant increase on their performances, compared to their results on the raw dataset. However, the preprocessing corroborated a significant increase, approximately 7%, on Linear Discriminant model. Its validation accuracy is 91.8% for raw dataset with 80%-20% split while it increases to 98.7%. Therefore, it can be deduced that preprocessing plays an important role when working with Linear Discriminant model. Apart from that, misclassification of Fe50, NiAl50 and Zn50 as Bi50 and Cu50 is common in all three models and may need further investigation on the issue.

Figures 8.54 to 8.56 show the best three models' test accuracies of preprocessed dataset with 80%-20% split.

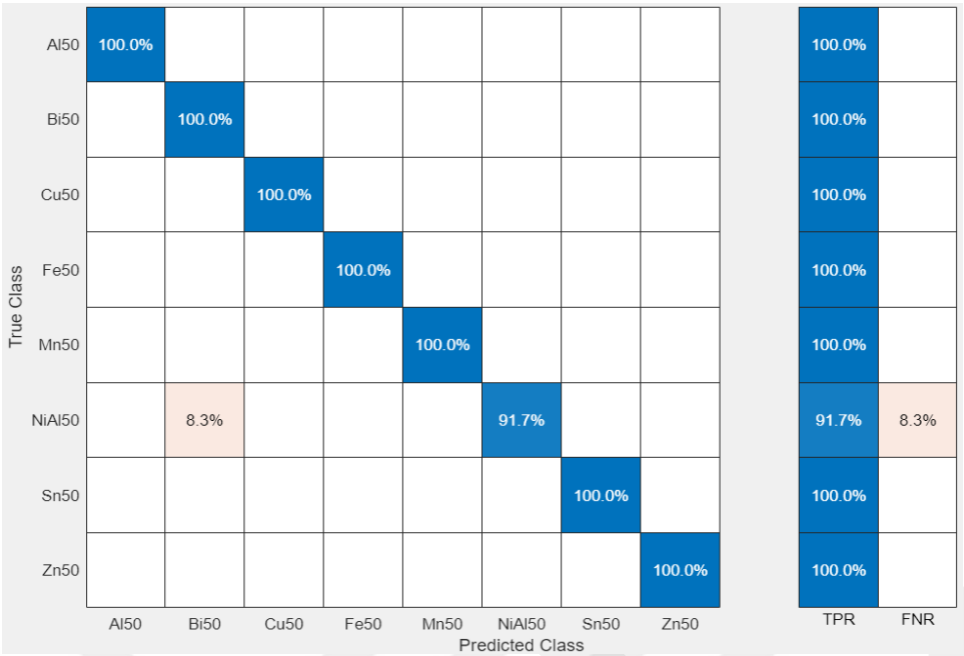

Figure 8.54 Quadratic SVM (99% Test Accuracy, Minerals Preprocessed Dataset, 80%-20% train test split)

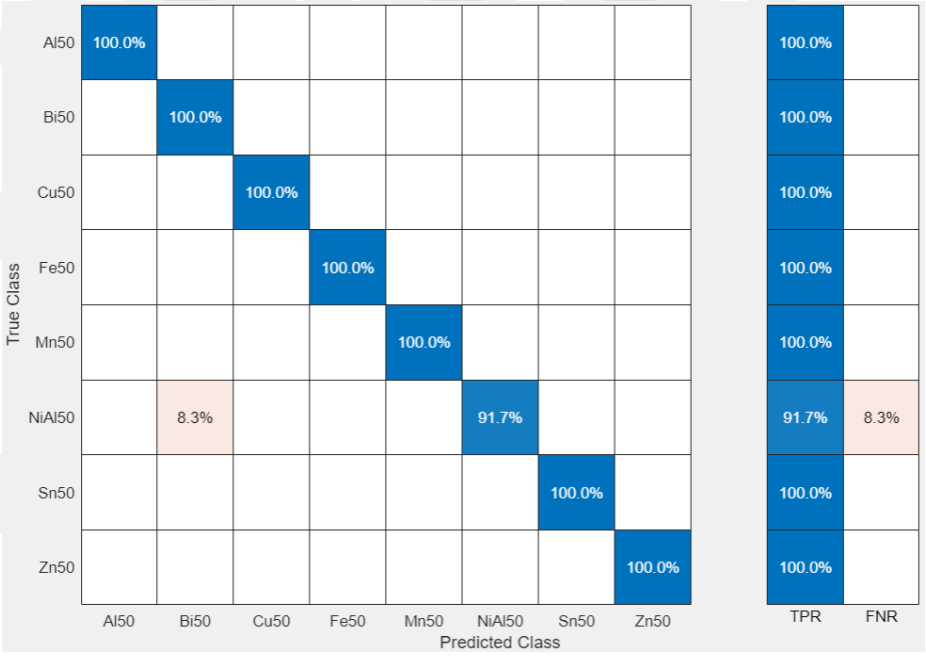

Figure 8.55 Cubic SVM (99% Test Accuracy, Minerals Preprocessed Dataset, 80%-20% train test split)

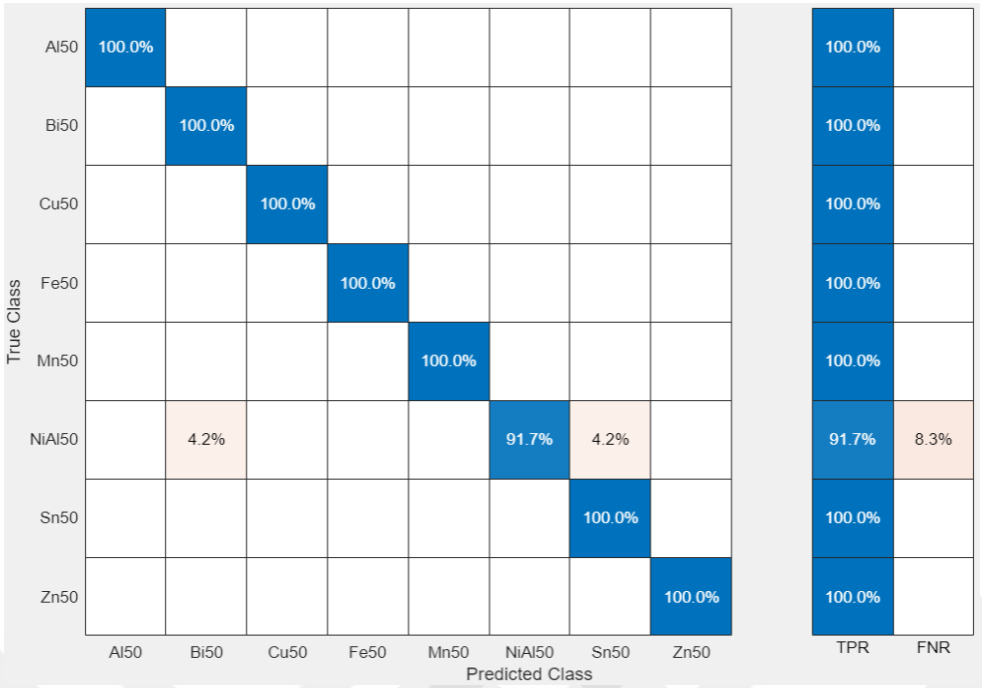

Figure 8.56 Medium Neural Network (99% Test Accuracy, Minerals Preprocessed Dataset, 80%-20% train test split)

From the above figures, the reprocessing did not manage to achieve a notable impact on the performances of the two SVM models. Yet, it increased the test performance of Medium NN from 93.8% to 99%. Additionally, it is observed that all three models consistently misclassified NiAl50 as Bi50. Likely to the previous repeating misclassification errors, NiAl50 and Bi50 might share a certain amount of similarity which prevented them to be distinguished with 100% accuracy. However, this misclassification remains low and other minerals were correctly classified.

CHAPTER 9

CONCLUSION

In this work, preprocessing methods and principal component analysis (PCA) combined supervised machine learning models were used together to classify two datasets; which contain paracetamol based two different pharmaceutical samples along with their different concentrations, and mineral samples having 8 different kinds of minerals.

In the pharmaceutical dataset, the best validation result was 84.4% using the Linear Discriminant and Ensemble of Subspace Discriminant classifier algorithms for an 80%-20% train/test split of the preprocessed data. The best test result, 87.5%, was obtained using the same models and split. Additionally, the 70%–30% split version of the preprocessed data used in the Linear SVM model had the same result for test accuracy. The three models stated above are promising, but more work has to be done on them. If the issues of noise, underfitting, and overfitting are resolved, the accuracy of all models can be increased. Thus, preprocessing may be used to PCA integrated with machine learning algorithms in a way that is more efficient, quicker, and simpler than previous approaches in the area.

In mineral dataset, the best validation result was acquired as 99.6% with Quadratic SVM classifier algorithms for 70%-30% train test split of the preprocessed data. The best test result is 99.0% and it is belong to both Quadratic SVM and Cubic SVM models with both of two splits. Additionally, it was observed that preprocessing did not make a major impact on the overall performance. Cubic and Quadratic SVM were usually among the best three models in every situation and their accuracies were satisfying even working on raw datasets. Thus, it can be commented that preprocessing work can be skipped in order to save time and resource when working with a dataset, expected to encounter the minerals in this study.

Finally, overall accuracy results of this study are lower than the results in the literature. The main reason behind these outcomes are the preparation of the samples and intense calibration of the laser system. In literature, samples had homogeneous

distribution and they were excited by intensely calibrated laser. However in this study, the samples have random distribution and they were excited with a laser system with default parameters, defined by its manufacturer.

REFERENCES

[1] N. Almsellati, O. Swesi, and S. Aladuli, "Quantitative analysis of pharmaceutical tablets using LIBS-technique", *International Journal of Development*, vol. 8, no. 1, pp. 1–7, 2019.

[2] S. Beldjilali, D. Borivent, L. Mercadier, E. Mothe, G. Clair, and J. Hermann, "Evaluation of minor element concentrations in potatoes using laser-induced breakdown spectroscopy" ,*Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 65, no. 8, pp. 727–733, 2010.

[3] G. Bilge, B. Sezer, I. H. Boyaci, K. E. Eseller, and H. Berberoglu, "Performance evaluation of laser induced breakdown spectroscopy in the measurement of liquid and solid samples", *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 145, pp. 115–121, 2018.

[4] G. Bilge, K. E. Eseller, H. Berberoglu, B. Sezer, U. Tamer, and I. H. Boyaci, "Comparison of different calibration techniques of laser induced breakdown spectroscopy in bakery products: On NaCl Measurement", *Journal of the European Optical Society-Rapid Publications*, vol. 17, no. 1, 2021.

[5] M. F. Bustamante, C. A. Rinaldi, and J. C. Ferrero, "Laser induced breakdown spectroscopy characterization of CA in a soil depth profile", *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 57, no. 2, pp. 303–309, 2002.

[6] D. A. Rusak, B. C. Castle, B. W. Smith, and J. D. Winefordner, "Fundamentals and applications of laser-induced breakdown spectroscopy", *Critical Reviews in Analytical Chemistry*, vol. 27, no. 4, pp. 257–290, 1997.

[7] D. Diaz, A. Molina, and D. W. Hahn, "Laser-induced breakdown spectroscopy and principal component analysis for the classification of spectra from gold-bearing ores", *Applied Spectroscopy*, vol. 74, no. 1, pp. 42–54, 2019.

[8] K.-B. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? an empirical study", *Multiple Classifier Systems*, pp. 278–285, 2005.

[9] J. Feng, Z. Wang, Z. Li, and W. Ni, “Study to reduce laser-induced breakdown spectroscopy measurement uncertainty using plasma characteristic parameters”, *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 65, no. 7, pp. 549–556, 2010.

[10] J. M. Gomba, C. D'Angelo, D. Bertuccelli, and G. Bertuccelli, “Spectroscopic characterization of Laser induced breakdown in aluminium–lithium alloy samples for quantitative determination of traces”, *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 56, no. 6, pp. 695–705, 2001.

[11] T. Hastie, J. Friedman, and R. Tisbshirani, *The elements of Statistical Learning: Data Mining, Inference, and prediction*, New York: Springer, 2017.

[12] T. Kim and C.-T. Li, “Laser-induced breakdown spectroscopy”, *Advanced Aspects of Spectroscopy,* 2012.

[13] W. B. Lee, J. Wu, Y. I. Lee, and J. Sneddon, “Recent applications of laser-induced breakdown spectrometry: A review of Material Approaches”, *Applied Spectroscopy Reviews*, vol. 39, no. 1, pp. 27–97, 2004.

[14] J. Li, J. Lu, Z. Lin, S. Gong, C. Xie, L. Chang, L. Yang, and P. Li, “Effects of experimental parameters on elemental analysis of coal by laser-induced breakdown spectroscopy”, *Optics & Laser Technology*, vol. 41, no. 8, pp. 907–913, 2009.

[15] P. Maravelaki-Kalaitzaki, D. Anglos, V. Kilikoglou, and V. Zafiropulos, “Compositional characterization of encrustation on marble with laser induced breakdown spectroscopy”, *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 56, no. 6, pp. 887–903, 2001.

[16] S. Pandhija and A. K. Rai, “Laser-induced breakdown spectroscopy: A versatile tool for monitoring traces in materials”, *Pramana*, vol. 70, no. 3, pp. 553–563, 2008.

[17] S. Pandhija, N. K. Rai, A. K. Rai, and S. N. Thakur, “Contaminant concentration in environmental samples using LIBS and CF-libs”, *Applied Physics B*, vol. 98, no. 1, pp. 231–241, 2009.

[18] J. Sneddon and Y.-I. Lee, “Novel and recent applications of elemental determination by laser-induced breakdown spectrometry”, *Analytical Letters*, vol. 32, no. 11, pp. 2143–2162, 1999.

[19] L. St-Onge, E. Kwong, M. Sabsabi, and E. B. Vadas, “Quantitative analysis of pharmaceutical products by laser-induced breakdown spectroscopy”, *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 57, no. 7, pp. 1131–1140, 2002.

[20] L. St-Onge, E. Kwong, M. Sabsabi, and E. B. Vadas, “Rapid analysis of liquid formulations containing sodium chloride using laser-induced breakdown spectroscopy”, *Journal of Pharmaceutical and Biomedical Analysis*, vol. 36, no. 2, pp. 277–284, 2004.

[21] E. Tognoni, V. Palleschi, M. Corsi, and G. Cristoforetti, “Quantitative micro-analysis by laser-induced breakdown spectroscopy: A review of the experimental approaches”, *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 57, no. 7, pp. 1115–1130, 2002.

[22] E. Tognoni and G. Cristoforetti, “Signal and noise in laser induced breakdown spectroscopy: An introductory review”, *Optics & Laser Technology*, vol. 79, pp. 164–172, 2016.

[23] J. Wang, X. Liao, P. Zheng, S. Xue, and R. Peng, “Classification of Chinese herbal medicine by laser-induced breakdown spectroscopy with principal component analysis and Artificial Neural Network”, *Analytical Letters*, vol. 51, no. 4, pp. 575–586, 2017.

[24] S. Wold, “Chemometrics; what do we mean with it, and what do we want from it?”, *Chemometrics and Intelligent Laboratory Systems*, vol. 30, no. 1, pp. 109–115, 1995.

[25] X. Zhang, F. Zhang, H.-te Kung, P. Shi, A. Yushanjiang, and S. Zhu, “Estimation of the Fe and Cu contents of the surface water in the Ebinur Lake basin based on Libs and a machine learning algorithm”, *International Journal of Environmental Research and Public Health*, vol. 15, no. 11, p. 2390, 2018.

[26] Y. Zhao, M. Lamine Guindo, X. Xu, M. Sun, J. Peng, F. Liu, and Y. He, “Deep learning associated with laser-induced breakdown spectroscopy (LIBS) for the

prediction of lead in soil", *Applied Spectroscopy*, vol. 73, no. 5, pp. 565–573, 2019.

[27] D. A. Cremers and L. J. Radziemski, *Handbook of Laser-induced breakdown spectroscopy*, West Sussex, UK, John Wiley & Sons, 2013.

[28] I. Schechter and A. W. Miziolek, *Laser-induced breakdown spectroscopy (LIBS)*, Cambridge, UK, Cambridge University Press, 2006.

[29] S. N. Thakur and J. P. Singh, "Fundamentals of laser induced breakdown spectroscopy", *Laser-Induced Breakdown Spectroscopy*, pp. 3–21, 2007.

[30] J. L. Gottfried, J. De Lucia, and Frank C., *Laser-induced breakdown spectroscopy: Capabilities and applications*, Aberdeen, USA, Army Research Lab Aberdeen Provıng Ground Md Weapons And Materials Research Directorate, 2010.

[31] Demtröder W., "Spectroscopic Instrumentation", *Laser spectroscopy*, Berlin: Springer, 2016, pp. 113–256.

[32] S. Xu, X. Sun, B. Zeng, W. Chu, J. Zhao, W. Liu, Y. Cheng, Z. Xu, and S. L. Chin, "Simple method of measuring laser peak intensity inside femtosecond laser filament in air" *Optics Express*, vol. 20, no. 1, p. 299, 2011.

[33] A. Talebpour, M. Abdel-Fattah, A. D. Bandrauk, and S. L. Chin, "Spectroscopy of the gases interacting with intense femtosecond laser pulses", *Laser physics*, 01-Jan-1970. [Online]. Available: http://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=994313. [Accessed: 01-Feb-2023].

[34] A. Talebpour, M. Abdel-Fattah, and S. L. Chin, "Focusing limits of intense ultrafast laser pulses in a high pressure gas: Road to new spectroscopic source", *Optics Communications*, vol. 183, no. 5-6, pp. 479–484, 2000.

[35] Y. E. Geints and A. A. Zemlyanov, "On the focusing limit of high-power femtosecond laser pulse propagation in air", *The European Physical Journal D*, vol. 55, no. 3, pp. 745–754, 2009.

[36] R. Noll, *Laser-induced breakdown spectroscopy fundamentals and applications*, Heidelberg: Springer-Verlag Berlin Heidelberg, 2012.

[37] F. Anabitarte, A. Cobo, and J. M. Lopez-Higuera, "Laser-induced breakdown spectroscopy: Fundamentals, applications, and challenges" ISRN *Spectroscopy*, vol. 2012, pp. 1–12, 2012.

[38] C. Pasquini, J. Cortez, L. M. Silva, and F. B. Gonzaga, "Laser induced breakdown spectroscopy", *Journal of the Brazilian Chemical Society*, vol. 18, no. 3, pp. 463–512, 2007.

[39] O. Samek, H. H. Telle, and D. C. S. Beddows, "Laser-induced breakdown spectroscopy: A tool for real-time, in vitro and in vivo identification of carious teeth", *BMC Oral Health*, vol. 1, no. 1, 2001.

[40] O. Samek, D. C. S. Beddows, H. H. Telle, G. W. Morris, M. Liska, and J. Kaiser, "Quantitative analysis of trace metal accumulation in teeth using laser-induced breakdown spectroscopy", *Applied Physics A Materials Science & Processing*, vol. 69, no. S1, 1999.

[41] O. Samek, M. Liska, J. Kaiser, D. C. S. Beddows, H. H. Telle, and S. V. Kukhlevsky, "Clinical application of laser-induced breakdown spectroscopy to the analysis of teeth and dental materials", *Journal of Clinical Laser Medicine & Surgery*, vol. 18, no. 6, pp. 281–289, 2000.

[42] L. Rokach and O. Maimon, "Data mining with decision trees", *Series in Machine Perception and Artificial Intelligence*, 2013.

[43] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning*, Cambridge, UK, Cambridge University Press, 2014.

[44] J. R. Quinlan, "Induction of Decision Trees" *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[45] L. Rokach and O. Maimon, "Top-down induction of decision trees classifiers—a survey", *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 4, pp. 476–487, 2005.

[46] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*, New York, USA, Routledge, 2017.

[47] C. Cortes and V. Vapnik, "Support-Vector Networks", *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[48] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[49] A. McCallum, "Graphical Models, Lecture2: Bayesian Network Representation." [Online]. Available: https://people.cs.umass.edu/~mccallum/courses/gm2011/02-bn-rep.pdf. [Accessed: 01-Feb-2023].

[50] S. M. Piryonesi and T. E. El-Diraby, "Role of data analytics in infrastructure asset management: Overcoming data size and quality problems", *Journal of Transportation Engineering*, Part B: Pavements, vol. 146, no. 2, p. 04020022, 2020.

[51] G. Brewka, "Artificial Intelligence—a modern approach by Stuart Russell and Peter Norvig, Prentice Hall. series in Artificial Intelligence, Englewood Cliffs, NJ.", *The Knowledge Engineering Review*, vol. 11, no. 1, pp. 78–79, 1996.

[52] E. Fix and J. L. Hodges, "Discriminatory analysis: Nonparametric discrimination: Consistency properties", *PsycEXTRA Dataset*, 1951.

[53] T. Cover and P. Hart, "Nearest neighbor Pattern Classification", *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[54] C. Naulak, "A comparative study of naive Bayes classifiers with improved technique on text classification", Internet: https://www.techrxiv.org/articles/preprint/A_comparative_study_of_Naive_BBaye_Classifiers_with_improved_technique_on_Text_Classification/19918361, May 31, 2022 , [Feb. 2, 2023]

[55] K. Nordhausen, "The elements of Statistical Learning: Data Mining, Inference, and prediction, second edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman" *International Statistical Review*, vol. 77, no. 3, pp. 482–482, 2009.

[56] D. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study”, *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 1999.

[57] R. Polikar, “Ensemble based systems in decision making”, *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[58] L. Rokach, “Ensemble-based classifiers”, *Artificial Intelligence Review*, vol. 33, no. 1-2, pp. 1–39, 2009.

[59] L. I. Kuncheva and C. J. Whitaker, “Measures of diversity in classifier ensembles”, *Machine Learning*, vol. 51, no. 2, pp. 181–207, 2003.

[60] P. Sollich and A. Krogh, “Learning with ensembles: How overfitting can be useful”, *Advances in Neural Information Processing Systems*, 01-Jan-1995. [Online]. Available: https://proceedings.neurips.cc/paper/1995/hash/1019c8091693ef5c5f55970346633f92-Abstract.html. [Accessed: 01-Feb-2023].

[61] G. Brown, J. Wyatt, R. Harris, and X. Yao, “Diversity creation methods: A survey and Categorisation”, *Information Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[62] J. J. Garcia Adeva, U. Cervino Beresi, and R. A. Calvo, “Accuracy and diversity in ensembles of text categorisers”, *CLEI Electronic Journal*, vol. 8, no. 2, 2005.

[63] T. K. Ho, "Random decision forests," *Proceedings of 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, 1995, pp. 278-282 vol.1.

[64] M. Gashler, C. Giraud-Carrier and T. Martinez, "Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous," *2008 Seventh International Conference on Machine Learning and Applications*, San Diego, CA, USA, 2008, pp. 900-905

[65] Y. Liu and X. Yao, “Ensemble learning via negative correlation”, *Neural Networks*, vol. 12, no. 10, pp. 1399–1404, 1999.

[66] R. Shoham and H. Permuter, “Amended cross-entropy cost: An approach for encouraging diversity in classification ensemble (brief announcement)”, *Lecture Notes in Computer Science*, pp. 202–207, 2019.

[67] Z. R. Yang and Z. Yang, “Artificial Neural Networks”, *Comprehensive Biomedical Physics*, pp. 1–17, 2014.

[68] L. H. Hardesty, “Explained: Neural network”, *MIT News*, Massachusetts Institute of Technology, 14-Apr-2017. [Online]. Available: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414. [Accessed: 01-Feb-2023].

[69] D. Pyle, *Data preparation for Data Mining*, San Francisco: Morgan Kaufmann, 2007.

[70] D. Chicco, “Ten quick tips for machine learning in Computational Biology - biodata mining”, *BioMed Central*, 08-Dec-2017. [Online]. Available: https://biodatamining.biomedcentral.com/articles/10.1186/s13040-017-0155-3. [Accessed: 01-Feb-2023].

[71] P. Oliveri, C. Malegori, R. Simonetti, and M. Casale, “The impact of signal pre-processing on the final interpretation of Analytical Outcomes – A tutorial”, *Analytica Chimica Acta*, vol. 1058, pp. 9–17, 2019.

[72] S. Wu, “A review on coarse warranty data and analysis”, *Reliability Engineering & System Safety*, vol. 114, pp. 1–11, 2013.

[73] “National Center for Education Statistics (NCES) home page, part of the U.S. Department of Education”, [Online]. Available: https://nces.ed.gov/. [Accessed: 01-Feb-2023].

[74] “UNECE”. [Online]. Available: https://unece.org/. [Accessed: 01-Feb-2023].

[75] S. C. Government of Canada, “3.4 processing 3.4.3 editing”, 02-Sep-2021. [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch3/editing-edition/5214781-eng.htm. [Accessed: 01-Feb-2023].

[76] “Statistics: Power from data!”, Government of Canada, Statistics Canada, 02-Sep-2021. [Online]. Available: https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-eng.htm. [Accessed: 01-Feb-2023].

[77] M. Li, “Evaluation of Travel Time Data Collection Techniques: A statistical analysis”, *International Journal of Traffic and Transportation Engineering*, vol. 2, no. 6, pp. 149–158, 2013.

[78] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Amsterdam: Elsevier/Morgan Kaufmann, 2012.

[79] Express Analytics, “What Is Data Wrangling? What are the steps in data wrangling?” *What Is Data Wrangling? Its Steps, Tools &* Techniques A Complete Guide, 30-Jul-2021.

Appendix A

PYTHON CODES

```
#Python Code for plotting Mineral Dataset

import seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

import plotly.express as px

from mpl_toolkits.mplot3d import Axes3D


#To plot other first 3 PCs of Mineral train datas; please try these: YTrain_pcscore_MINER_08_NOSTAND, YTrain_pcscore_MINER_07_WITHSTAND, #and YTrain_pcscore_MINER_07_NOSTAND

dfmin = pd.read_csv('YTrain_pcscore_MINER_08_WITHSTAND.csv', delimiter=',')

dfmin.head(3)

dfmin.info()


dfmin.rename(columns={'pcscoremineralsingle1': 'PCA_1', 'pcscoremineralsingle2': 'PCA_2', 'pcscoremineralsingle3': 'PCA_3'}, inplace=True)


fig = px.scatter_3d(dfmin, x='PCA_1', y='PCA_2', z='PCA_3',

              color='Minerals',

              title="3D Scatter Plot",width=1200, height=1200)

fig.show()
```

```
#Python Code for plotting Medicines Dataset
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import plotly.express as px

# To plot other first 3 PCs of Medicines train datas; please try these: YTrain_pcscore_MED_08_NOSTAND, YTrain_pcscore_MED_07_WITHSTAND,
# and YTrain_pcscore_MED_07_NOSTAND
dfmed = pd.read_csv('YTrain_pcscore_MED_08_WITHSTAND.csv', delimiter=',')
dfmed.head(3)

dfmed.info()

dfmed.rename(columns={'pcscoremedsingle1': 'PCA_1', 'pcscoremedsingle2': 'PCA_2', 'pcscoremedsingle3': 'PCA_3'}, inplace=True)

fig = px.scatter_3d(dfmed, x='PCA_1', y='PCA_2', z='PCA_3',
              color='Medicines',
              title="3D Scatter Plot",width=1000, height=1000)
fig.show()
```

Appendix B

MATLAB CODES

```
% This code is used for preprocessing and plotting scatter graphics for
% medicines dataset

% CLEARING
clc
clear all


datamed=readtable("remodified_ilac.csv",'ReadVariableNames',true,'Varia
bleNamingRule','preserve',"TextType","char");


datamedresp=datamed(:,1);
datamedpred=datamed(:,2:end);

%preprocessing
datamedpred=standardizeMissing(datamedpred,{Inf,'N/A'});
datamedpred=fillmissing(datamedpred,'movmean',5);
datamedpred=filloutliers(datamedpred,'pchip','quartiles');
datamedpred=smoothdata(datamedpred,"sgolay");
datamedpred=normalize(datamedpred,'range',[0,1]);

%uniting
datamedstand=[datamedresp datamedpred];
datamedstand.Properties.VariableNames=datamed.Properties.VariableNames;


%partitioning data
%train and test ratios can be arranged differently.
cv=cvpartition(datamedstand.Medicines,'HoldOut',0.20);
Train=datamedstand(training(cv),:);
Test=datamedstand(test(cv),:);


XTrain=table2array(Train08IlacWithStand(:,2:end));
YTrain=Train08IlacWithStand.Medicines;
XTest=table2array(Test08IlacWithStand(:,2:end));
YTest=Test08IlacWithStand.Medicines;


[coeffmed,pcscoremed,~,~,explainedmed,mumed] = pca(XTrain);
explainedmed

pcscoremedsingle=single(pcscoremed);
pcscoremedtable=array2table(pcscoremedsingle);
```

```
YTrain_pcscoremed=[YTrain pcscoremedtable];
YTrain_pcscoremed.Properties.VariableNames{'Var1'} = 'Medicines';


writetable(YTrain_pcscoremed,'C:\Users\MONSTER\Desktop\GUNCEL_YL_TEZ_PC
A_IMPROVE_VISUAL\YTrain_pcscore_MED_08_WITHSTAND.csv','Delimiter',',')


gscatter(pcscoremed(:,1),pcscoremed(:,2),YTrain)
axis equal
xlabel('1st Principal Component')
ylabel('2nd Principal Component')


gscatter(pcscoremed(:,1),pcscoremed(:,3),YTrain)
axis equal
xlabel('1st Principal Component')
ylabel('3rd Principal Component')


gscatter(pcscoremed(:,2),pcscoremed(:,3),YTrain)
axis equal
xlabel('2nd Principal Component')
ylabel('3rd Principal Component')


scatter3(pcscoremed(:,1),pcscoremed(:,2),pcscoremed(:,3))
axis equal
xlabel('1st Principal Component')
ylabel('2nd Principal Component')
zlabel('3rd Principal Component')
```

```
% This code is used for preprocessing and plotting scatter graphics for
% minerals dataset
% CLEARING
clc
clear all


datamineralstore=datastore("DATAMINERALS",'IncludeSubfolders',true,'Rea
dVariableNames',true,'VariableNamingRule','preserve',"TextType","char")
;
datamineralraw=readall(datamineralstore);



writetable(datamineralraw,'D:\YUKSEK_LISANS_TEZ\MINERALS_PCA_APPS_CLASS
IFICATION_LEARNERS_RESULTS\datamineralraw.csv','Delimiter',',')
%preprocessing
datamineralresp=datamineralraw(:,1);
datamineralpred=datamineralraw(:,2:end);
datamineralpred=standardizeMissing(datamineralpred,{Inf,'N/A'});
datamineralpred=fillmissing(datamineralpred,'pchip');
datamineralpred=filloutliers(datamineralpred,'pchip','quartiles');
datamineralpred=smoothdata(datamineralpred,"sgolay");
datamineralpred=normalize(datamineralpred,'range',[0,1]);

%uniting
datamineralstand=[datamineralresp datamineralpred];
datamineralstand.Properties.VariableNames=datamineralraw.Properties.


%partitioning data
%train and test ratios can be arranged differently.
cv=cvpartition(datamineralstand.Minerals,'HoldOut',0.20);
Train=datamineralstand(training(cv),:);
Test=datamineralstand(test(cv),:);


XTrain=table2array(Train08MineralWithStand(:,2:end));
YTrain=Train08MineralWithStand.Minerals;
XTest=table2array(Test08MineralWithStand(:,2:end));
YTest=Test08MineralWithStand.Minerals;


[coeffmineral,pcscoremineral,~,~,explainedmineral,mumineral] =
pca(XTrain);
explainedmineral

pcscoremineralsingle=single(pcscoremineral);
pcscoremineraltable=array2table(pcscoremineralsingle);
```

```
YTrain_pcscoremineral=[YTrain pcscoremineraltable];
YTrain_pcscoremineral.Properties.VariableNames{'Var1'} = 'Minerals';


writetable(YTrain_pcscoremineral,'C:\Users\MONSTER\Desktop\GUNCEL_YL_TEZ_PCA_IMPROVE_VISUAL\YTrain_pcscoremineral_MINER_08_WITHSTAND.csv','Delimiter',',')


gscatter(pcscoremineral(:,1),pcscoremineral(:,2),YTrain)
axis equal
xlabel('1st Principal Component')
ylabel('2nd Principal Component')


gscatter(pcscoremineral(:,1),pcscoremineral(:,3),YTrain)
axis equal
xlabel('1st Principal Component')
ylabel('3rd Principal Component')


gscatter(pcscoremineral(:,2),pcscoremineral(:,3),YTrain)
axis equal
xlabel('2nd Principal Component')
ylabel('3rd Principal Component')


scatter3(pcscoremineral(:,1),pcscoremineral(:,2),pcscoremineral(:,3))
axis equal
xlabel('1st Principal Component')
ylabel('2nd Principal Component')
zlabel('3rd Principal Component')
```