# Review on Effective Disease Prediction through Data Mining Techniques

Muhammad Nabeel[1], Shumaila Majeed[2], Mazhar Javed Awan[1*], Hooria Muslih-ud-Din[3], Mashal Wasique[2] and Rabia Nasir[4]

[1]Department of Software Engineering, University of Management and Technology, Lahore, Pakistan
[2]Department of Computer Science, Fatima Jinnah Women University, Rawalpindi, Pakistan.
[3]Department of Computer Science, National University of Computer and Emerging Sciences, Lahore, Pakistan.
[4]Department of Anesthesia, Lahore General Hospital, Lahore, Pakistan.
Corresponding author: *mazhar.awan@umt.edu.pk

*Abstract:* Hidden and unknown pattern are extracted from large data sets by performing several combinations of techniques from database and machine learning. Data mining plays a significant role for handling a huge amount of data. Data mining deals with heterogeneity, privacy and correctness of data. Moreover, medical data mining is tremendously important research area and significant attempts are made in this area in recent years because inaccuracy in medical data systems may cause seriously disingenuous medical treatments. Medical data sets should be analyzed using suitable mining algorithms. To perform related operations, techniques of data mining have been used in developing medical systems for prediction of diseases through a set of medical data set. This paper reviews state of the art data mining algorithms for predicting different diseases and to analyze the performance of classification techniques i.e. Naive Bayes (NB), J48, REF Tree, Sequential Minimal Optimization (SMO), Multi-Layer Perceptron and Vote on different data sets of different diseases i.e. chronic kidney disease (CKD), heart disease, liver and diabetes. The experimental setup for performance evaluation of various algorithms using disease data sets retrieved from UCI respiratory has been made in WEKA tool. Values of different parameters i.e. correctly classified instances, precision, recall and F-Measure, time taken are analyzed by applying different classification algorithms.

*Keywords*: Naive Bayes, J48; REFTree, Sequential Minimal Optimization (SMO), multi-layer perceptron, vote, classification, prediction

## 1. Introduction

Data mining is the technique of knowledge discovery in which the knowledge is gathered by examining the data which might be hidden in extremely large sources, these sources are analyzed from several perspectives using different techniques and then the extracted information is summarized into useful information. It is a process in which information from past records are extracted for making important decisions for future predictions. Data mining techniques are becoming an important area of research for effective analysis of large data as the complexity and size of the data increases [1]. Data mining is used in many domains i.e. image mining, opinion mining, web mining, text mining, graph mining, medical data systems. It has become an important medical research area for finding unknown patterns in medical data. Medical professionals can examine the diseases on prediction analysis given by prediction model. In medical field data mining technique plays a vital role to predict different diseases. In many cases doctors may not be able to predict whether patient is suffering from one or more diseases at the same time. With the advent of new developments in the field of medication, a lot of data about different diseases have been gathered and are accessible to the research community [2]. Many challenges and opportunities are faced, according to the data mining perspective, mining big data has opened many new challenges and opportunities. A small number of data mining applications have been effectively provoked in different areas like extortion identification, retail, astronomy, social insurance, social media, money, banking, media transmission, climate modeling, medical, telecommunication, and hazard analysis etc. are not many to name [3]. In healthcare huge amount of data is being generated so, processing and analysis is required for knowledge extraction from

Received: December 28th, 2020. Accepted: September 30th, 2021
DOI: 10.15676/ijeei.2021.13.3.13

such huge data. Mining algorithms predict the disease of patients using suitable learning strategy. Diseases like chronic kidney disease (CKD), hepatitis, cancer disease and diabetes have become a worldwide health issue and therefore prediction of such type of diseases is the concerned area for researchers. Our work mainly focuses on analyzing classification algorithms like Naive Bayes (NB) and Artificial Neural Network (ANN), J48, REF Tree for different life threatening diseases like CKD. Mining techniques used in health care are described in fig1. There are two main categories of data mining known as supervised and unsupervised [4]. Both approaches have different applications and efficiency for analyzing and predicting the diseases. These techniques mentioned above are used in medical field accordingly to predict diseases and for making decision for treatment of patients. Classification is a supervised technique in which objects are assigned in a collection to target classes. Decision tree, ANN, SVM, NB, etc are approaches of classification. Different approaches are used for different purposes in healthcare. In clustering similar types of objects are categorized in the same group. K-means, K- medoids, agglomerative, divisive, DBSCAN etc are some of the techniques of clustering. Association is the possibility of occurrence of objects in a set. Further classification of Apriori is association. Hierarchy of data mining is shown in figure 1.

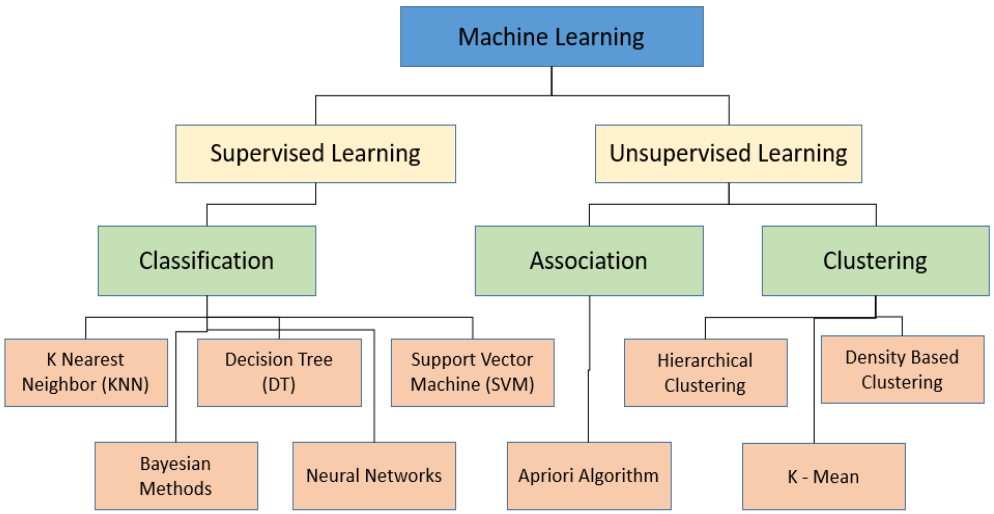



Figure 1. The Taxonomy of Data Mining

So far, review of many mining techniques are done for different diseases separately but comparative analysis of all these algorithms on different diseases was not done by any other study. This study presents comparative survey of different algorithms on four different diseases as well as the algorithms are implemented in Weka and discussed in terms of parameters i.e. correctly classified instances, precision, recall and F-Measure, time taken. Further organization of paper in different sections are given as, section II gives a systematic review of algorithms used for predicting different diseases. Section III contains the simulation results and discussion.

## 1. Literature Review

Data mining in healthcare is one of the most challenging field of application for knowledge discovery. Medical data mining is challenging because of complexity, diversity and huge amount of data set. As the provided data set in health care are heterogeneous in nature, so predicting disease of such heterogeneous and fragmented data is a challenging task [5]. This section provides a comprehensive literature review of various algorithms of data mining used for predicting different diseases. The performance analysis of various algorithms for prediction of different diseases given in literature are shown in table 1, table2, table 3 and table 4 respectively.

*A. Heart Diseases*

Naive Bayes algorithm is used by [13] for predicting heart disease accurately. Neural Network (NN), Decision Tree (DT) and NB are compared in the study and Naive Bayes performs the best. Algorithms such as C5.0, SVM, KNN, NN and Logistic Regression are discussed by [25]. C5.0 Decision tree has the highest accuracy of 93.02% and thus improved the prediction system. Issues in predicting the heart disease are presented by [8] and algorithms such as KStar, J48, SMO, NaiveBayes and Multilayer Perceptron are investigated and implemented using tool Weka SMO shows the highest accuracy of 84.074%. KNN, Neural Network, Decision Tree Algorithm, Naive Bayes are various classification techniques analyzed by [26] and it is concluded that KNN and ID3 gives the maximum accuracy of 80.6% in heart disease risk prediction [27]. The comprehensive review of prediction algorithms for heart disease are described in table1.

Table 1. Comprehensive review of Mining Techniques for Heart Diseases

| Paper | Models | Highest Accuracy Model | Accuracy |
|---|---|---|---|
| [13] | Naive Bayes, Neural Network and Decision Tree. | Naive Bayes | 99% |
| [25] | C5.0, Neural Network,SVM, KNN and Logistic Regression, | C5.0 Decision tree | 93.0% |
| [8] | Custer, J48, SMO, Bayes Net, Multi-layer Perceptron | SMO | 84.07% |
| [26] | NB, ID3, KNN, Decision Tree, Neural Network. | KNN and ID3 | 80.6 |
| [27] | NB, Discriminant, Decision Tree, Random Forest, and SVM | Decision tree and Random Forest | 99.0% |
| [7] | Naive Bayes, Random Forest | Naive Bayes | 84.15%, 84.16% |
| [28] | Apriori , SVM, Fuzzy rule based, Neural Network, Regression Trees, Fuzzy Clinical decision support system, KNN, genetic algorithm, type-2 Fuzzy Logic, Decision Tree | SVM | N/A |
| [14] | SVM combined with genetic algorithm, C5.0, Na¨ıve Bayesian, J4.8, KNN, Neural Network, decision tree and Fuzzy | SVM | N/A |
| [9] | NB , KNN, Decision Tree, Logistic Regression | Logistic Regression | N/A |
| [6] | SVM, KNN, ANN | SVM | 85% |
| [29] | NB, Decision Trees i.e. C5.0 boosted, CART, C4.5, and C5.0. random forest algorithms | CART | 87% |
| [5] | Naive Bayes, KNN, Decision Tree, SVM, Logistic Regression, Neural Network and Vote (a hybrid technique with NB and LR). | Vote | 87.4% |

The Random Forest, NB, Decision Tree, Discriminant and SVM for predicting the heart disease risk using two data set in MATLAB and concluded that Decision Tree, Random Forest gives the accuracy of 99.0 are best among all discussed algorithms [7].The purposes a hybrid algorithm of Random Forest and Naive Bayes and [7] concluded that the hybrid approach improves the prediction of heart Disease. Naïve Bayes obtain accuracy of 84.1584%, while Random Forest attained 84.1604%. [28] Algorithms of data mining i.e. Apriori algorithm, Support Vector Machine, Neural Network, Classification and Regression Trees, Fuzzy rule-based clinical decision support system, K-nearest neighbor, Genetic algorithm, Type-2 fuzzy

logic system, Decision Tree are discussed and it is concluded that SVM algorithm achieved best accuracy. Several analytic techniques i.e. SVM classifier with Genetic algorithm, Naive Bayesian, C5.0, Neural Network, KNN, J48, Decision tree and Fuzzy on FHS (Framingham Heart Study) are analyzed by [14] and concluded that SVM classifier with genetic algorithm outperforms. Naive Bayes, Decision Tree, KNN are analyzed by [9] for heart disease prediction. [6] Discussed classification techniques such as KNN, SVM and ANN for heart disease prediction. Data sets used for simulation are C-level and standard data set obtained from UCI machine learning repository. Results of the study shows that SVM perform better than that of other algorithms with accuracy of 85.1852%. The experiments are done using MATLAB. [29]Analyzed algorithms Naive Bayes classifier, decision trees (C4.5, CART, C5.0 boosted, and C5.0) and Random Forest for prediction of acute rheumatic fever on cardiac disease and conclusion of the study shows that CART outperforms by 87%. [5]Studied the significant features of algorithms such as KNN, Decision Tree, SVM, NB, Neural Network and Logistic Regression, Vote. Study concluded that Vote gives the best performance and achieved the 87.4% accuracy.

*B. Diabetes Diseases*

ANN, K-fold cross validation, Vector support machine, KNN method algorithms are discussed and analyzed by [17] and study concluded that SVM outperforms with an accuracy of 81.77% to predict diabetes. Algorithms i.e. Naive Bayes Classifier, C4.5, SVM, KNN is reviewed by [16] using PIMA Indian diabetes data sets and concluded that C4.5 algorithm is more accurate than KNN with an accuracy of 78.25%. Bagging ensemble and Adaboost techniques along with J48, c4.5, decision tree using Canadian Primary Care database and Adaboost have more accuracy [30]. Data mining algorithms such as Random Forest, SOM also C4.5 are experimented on data set of Ministry of National Guard Health Affairs (MNGHA) on adult population by [31]. Result shows that Random Forest achieve highest accuracy of 90%. KNN, Naive Bayes, Decision Tree are implemented by [32] on PID data set to predict Diabetes mellitus. Analysis results shows that Decision Tree has attained best accuracy 75.65% as compared to other discussed algorithms. Simple CART and NB, J48 are compared by [33] and it is concluded that J48 and Simple CART are cost efficient and gives 99% result. SVM is better according to [14] for diabetes as compared to Bayesian Classifier, SVM, Neural Network and Decision Tree algorithm. Genetic algorithm, H-means plus clustering and EM algorithm, Random Forest Classifier are analyzed by [34] and results concluded that Random Forest results are best in terms of accuracy. Algorithms such as Bayesian, SVM, and Decision Tree are compared by [35] and a new hybrid technique is also proposed which attained 94% accuracy. Comprehensive review of techniques for Diabetes prediction are given in table 2.

*C. Liver Diseases*

Fuzzy based classification for Liver disorder prediction is applied by [23] and the purposed algorithm gives the accuracy of 94%. Liver diseases are predicted using classification algorithms Naive Bayes and SVM. Results shows that SVM performs better in predicting liver disease. Experiment is done using Matlab [21]. Decision Tree, NB, C4.5, SVM, Regression Tree and Back Propagation are some algorithms analyzed by [24] and results show that C4.5 performs better as compared to other algorithms. Performance of different algorithms like C5.0 and CHAID are compared in [36] using UCI repository data set.

Table 2. Comprehensive review of Mining Techniques for Diabetes Diseases

| Paper | Models | Highest Accuracy Model | Accuracy |
|---|---|---|---|
| [17] | K fold cross validation, CKNN, classification methods, SVM, LDA SVM ,Feed Forward Neural Network, Back propagation, ANN, Statistical Normalization. | SVM | 81.77% |
| [16] | Naive Bayes, SVM, C4.5, KNN. | C4.5 | 78.25% |
| [30] | Bagging ensemble , adaboost using J48 and C4.5. | Adaboost ensemble method | N/A |
| [31] | Self-organizing Map (SOM), C4.5 and Random Forest, | Random Forest | 90% |
| [32] | Decision Tree, NB, KNN, | Decision Tree | 75.65% |
| [33] | J48, CART and Naive bayes | J48 and CART is cost efficient | 99 % |
| [14] | Bayesian Classifier, Neural Network, Decision Tree and SVM algorithm. | SVM | N/A |
| [34] | EM algorithm, Genetic Algorithm, H-means+ clustering and Random Forest Classifier | Random Forest | N/A |
| [35] | Bayesian, SVM, Decision Tree, | Hybrid proposed method of decision tree and SVM | 94% |

By doing performance evaluation it is concluded that C5.0 algorithm using Boosting technique achieved 93.75% accuracy in predicting liver diseases.

Table 3. Comprehensive review of Mining Techniques for Liver Diseases

| Paper | Models | Highest Accuracy Model | Accuracy |
|---|---|---|---|
| [18] | ANN, SVM | ANN | 87.70 |
| [20] | J48, NB, SVM, Multilayer Perceptron, Decision Tree, Conjunctive Rule. | Multilayer Perceptron | 99.75% |
| [43] | Naive Bayes and SVM. | SVM | 76.32 |
| [19] | Decision Tree (C4.5), SVM and Bayesian Network | C4.5 | N/A |
| [22] | Decision Tree, Naive Bayes, KNN, Rule based , Back Propagation , SVM | Multilayer Perceptron, SVM, Radial Basis Function Naive Bayes, Random forest. | N/A |
| [44] | Back Propagation Neural Network, One Rule classifier, NB, Decision Table, Decision trees and KNN | Naive Bayes | 99.36 % |
| [45] | SVM, Decision Tree, NB and Linear Regression, Neural Network. | Neural Network | N/A |
| [46] | SVM,J48, Na¨ıve Bayes | SVM gave | 75.75% |
| [47] | K-Star, SVM, NB and J48 | J48 | 99% |
| [48] | multilayer perceptron, naive bayes and J48 decision tree | J48 | 87.3 |
| [49] | Naive Bayes, SVM, ANN and ANFIS | Naive Bayes and KNN | N/A |

Boosted C5.0 combined with Genetic Algorithm (GA) is proposed by [37] and concluded that accuracy is improved from 81 to 93%. Using WEKA [38] experimented Grading learning algorithms, logitboost, Bagging, Adaboost are applied to this data set and concluded that Grading algorithm have shown highest accuracy of 71.3551%. [39] Analyzed Naive Bayes (NB), C5.0,

K-means, KNN, Random Forest, C5.0 with Adaptive Boosting and concluded that C5.0 with Adaptive Boosting performs better. K-Nearest Neighbour outperforms among other algorithms discussed in the study such as Auto Neural, Random Forest and Logistic regression. Conclusion of the study have shown that KNN have highest accuracy of 99.794% [40]. Boosting C5.0 has the highest accuracy as compared to SVM, Exhaustive CHAID and CHAID according to [41]. Random forest (RF), ANN, NB and Logistic Regression are discussed in the study according to study Random Forest outperforms and attained accuracy of 87.48%. Comprehensive Review of liver Disease prediction algorithms is given in table 3.

*D. Kidney Diseases*

ANN and SVM are compared by [18] and concluded that ANN achieved better accuracy while SVM achieved better execution time. ANN shows 87.70% accuracy. [20]Multi-layer perceptron algorithm performs best among Naive Bayes, Decision Tree, SVM, Multilayer Perceptron, J48, conjunctive rule. Multi-layer perceptron attained 99.75% accuracy. The SVM and NB are compared by [44] and concluded that SVM outperform by achieving 76.32% accuracy. The SVM, C4.5, Decision Tree and Bayesian Network are Compared and discussed in [19].

Table 4. Comprehensive review of Mining Techniques for Kidney Diseases

| Paper | Models | Highest Accuracy Model | Accuracy |
|---|---|---|---|
| [18] | ANN, SVM | ANN | 87.70 |
| [20] | J48, NB, SVM, Multilayer Perceptron, Decision Tree, Conjunctive Rule. | Multilayer Perceptron | 99.75% |
| [43] | Naive Bayes and SVM. | SVM | 76.32 |
| [19] | Decision Tree (C4.5), SVM and Bayesian Network | C4.5 | N/A |
| [22] | Decision Tree, Naive Bayes, KNN, Rule based, Back Propagation , SVM | Multilayer Perceptron, SVM, Radial Basis Function Naive Bayes, Random forest. | N/A |
| [44] | Back Propagation Neural Network, One Rule classifier, NB, Decision Table, Decision trees and KNN | Naive Bayes | 99.36 % |
| [45] | SVM, Decision Tree, NB and Linear Regression, Neural Network. | Neural Network | N/A |
| [46] | SVM,J48, Naive Bayes | SVM gave | 75.75% |
| [47] | K-Star, SVM, NB and J48 | J48 | 99% |
| [48] | multilayer perceptron, naive bayes and J48 decision tree | J48 | 87.3 |
| [49] | Naive Bayes, SVM, ANN and ANFIS | Naive Bayes and KNN | N/A |

According to [19] C4.5 achieved better accuracy and execution time. Hidden pattern Relationship of CKD is discovered by [22] using classification techniques i.e. KNN, Decision Tree, ANN. For prediction of chronic kidney it is analyzed that Random Forest, SVM, NB, Multilayer Perceptron, Radial Basis and KNN Function enhance the accuracy. Algorithms such as Decision Table, NB, Decision trees, Back Propagation Neural Network, KNN and One Rule Classifier are compared by [45] for chronic kidney disease prediction and concluded that Naive Bayes gives best accuracy 99.36 %. Decision Tree, Linear Regressing, SVM, Neural Network and Naive Bayesian are experimented using UCI data base. Neural Network performs best for

chronic kidney disease prediction according to review by [46]. Techniques i.e. Artificial Neural Network, Decision Tree and KNN are implemented by [47] to discover hidden patterns of CKD, SVM, J48, and Naive Bayes SVM gave maximum accuracy of 75.75%. [48] K-Star, NB, SVM and J48 classifiers are some algorithms discussed in the study. Outcome of the study shows that J48 outperforms with 99%. [49-56] Implemented Multi-layer perceptron, NB and J48 decision tree was using WEKA in this study. J48 decision tree performs best 87.3%. Techniques like Naive Bayes, SVM, ANN and ANFIS Naive Bayes and KNN are compared by [57-62]. The study concluded that Naive Bayes and KNN performs best. Comprehensive review of algorithms for kidney prediction is shown in table 4.

## 2. Result Comparison

Heart can be affected by a different condition which includes blood vessel disease known as coronary artery disease (CAD), congenital heart defects, heart valve disease, heart infection and heart muscle. Diabetes occurs when the body does not process and use glucose from the food. Characteristics of diabetes include presence of autoantibodies, injury to pancreas, family history, physical stress, being overweight, high blood pressure, smoker, having low cholesterol and physically inactive. Common symptoms of liver disease include liver enlargement, portal hypertension, abnormal bleeding, severe itching, extreme tiredness, yellowing of the skin and eyes. Kidney is an important organ of the human body located at the lower back. Different characteristics of kidney failure are reduced urine, swelling of legs, persistent nausea, shortness of breath.

Results of classification techniques i.e. Naive Bayes, J48, REFTree, SMO, Multi-Layer Perceptron, Vote on different data sets of different diseases are discussed in this section.

### *A. Result of Heart Diseases*

For heart disease classification analysis, the data set is taken from UCI repository. Total number of instances is 270 and number of features is 14. It consists of both nominal and numerical data.

Table 5. Accuracy Evaluation of Data Mining and Machine Learning Algorithm on Heart Disease

| Algorithm | Correctly Classified Instances | Incorrectly Classified Instances | Kappa statistic | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|
| NaiveBayes | 226 | 44 | 0.6683 | 0.837 | 0.837 | 0.837 |
| Multilayer Perception | 211 | 59 | 0.5601 | 0.784 | 0.781 | 0.782 |
| SMO | 227 | 43 | 0.6762 | 0.841 | 0.841 | 0.840 |
| Random Forest | 229 | 41 | 0.6907 | 0.848 | 0.848 | 0.848 |
| Vote | 150 | 120 | 0 | 0.566 | 0.566 | 0.566 |
| Decision Table | 220 | 50 | 0.6244 | 0.815 | 0.815 | 0.815 |
| J48 | 207 | 63 | 0.5271 | 0.766 | 0.767 | 0.767 |

Characteristics of heart disease in dataset consist of age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) colored by flourosopy, thal: 3 = normal; 6 = fixed defect; 7 = reversable defect. The classification algorithm which performed best for heart

disease is random forest. Its classification accuracy is 83.77%. The simulation results are shown in table 5 while its graphical representation is shown in figure 2.

Random forest algorithm utilizes ensemble learning to solve complex problems. Random forest works like decision trees but it contains multiple decision trees which works well on characteristic of heart disease. RF machine learning algorithm linked outcome based on prediction of decision trees. F measure of random forest is around 83.77% which outperform as shown in the table. In Table 5 after random forest naivebayes algorithm gives better performance.

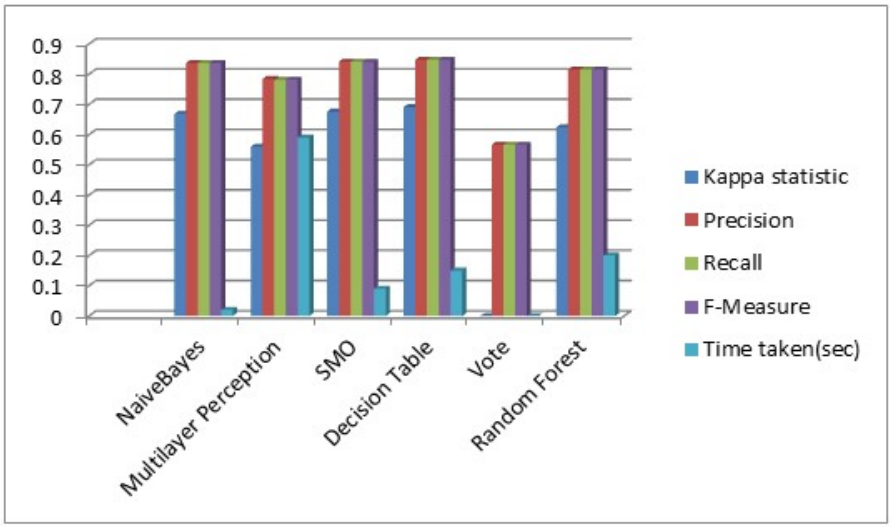



Figure 2. Bar graph of comparisons of heart diseases evaluation

*B. Result of Diabetes Diseases*

For diabetes disease classification analysis, the data set is taken from UCI repository. Total number of instances is 768. Data set contains of both nominal and numerical data. The classification algorithm SMO performed best in terms of correctly classified instances. Its classification accuracy is 77.34%. The simulation results are shown in figure 3.

Characteristics of diabetes includes Regular insulin dose, Unspecified special event, NPH insulin dose, UltraLente insulin dose, Unspecified blood glucose measurement, Typical exercise activity, Unspecified blood glucose measurement, Pre-breakfast blood glucose measurement, Post-breakfast blood glucose measurement, Pre-lunch blood glucose measurement, Post-lunch blood glucose measurement, Pre-supper blood glucose measurement, More-than-usual exercise activity, Post-supper blood glucose measurement, Pre-snack blood glucose measurement, Hypoglycemic symptoms, More-than-usual meal ingestion, Typical meal ingestion, Less-than-usual meal ingestion, Less-than-usual exercise activity.

Sequential minimal optimization (SMO) is an algorithm which solve quadratic problem during the training of support vector machines. On the data characteristics of diabetes which is discussed above SMO perform better than other algorithms. Table 6 explain the detail of evaluation model as shown in table SMO precision is 0.769 and recall is 0.773 overall F-measure is 0.750 which is better than other algorithm. NaiveBayes provides close accuracy still SMO outperform naivebayes.

Table 6. Evaluation of Data Mining and Machine Learning Algorithm on Diabetes

| Algorithm | Correctly Classified Instances | Incorrectly Classified Instances | Kappa statistic | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|
| NaiveBayes | 586 | 182 | 0.4664 | 0.759 | 0.763 | 0.760 |
| Multilayer Perception | 579 | 189 | 0.4484 | 0.750 | 0.754 | 0.751 |
| SMO | 594 | 174 | 0.4682 | 0.769 | 0.773 | 0.763 |
| Decision Table | 547 | 221 | 0.3492 | 0.706 | 0.712 | 0.708 |
| Vote | 500 | 268 | 0 | 0.651 | 0.651 | 0.789 |
| Random Forest | 578 | 190 | 0.4459 | 0.749 | 0.753 | 0.750 |
| J48 | 567 | 201 | 0.4146 | 0.735 | 0.738 | 0.736 |

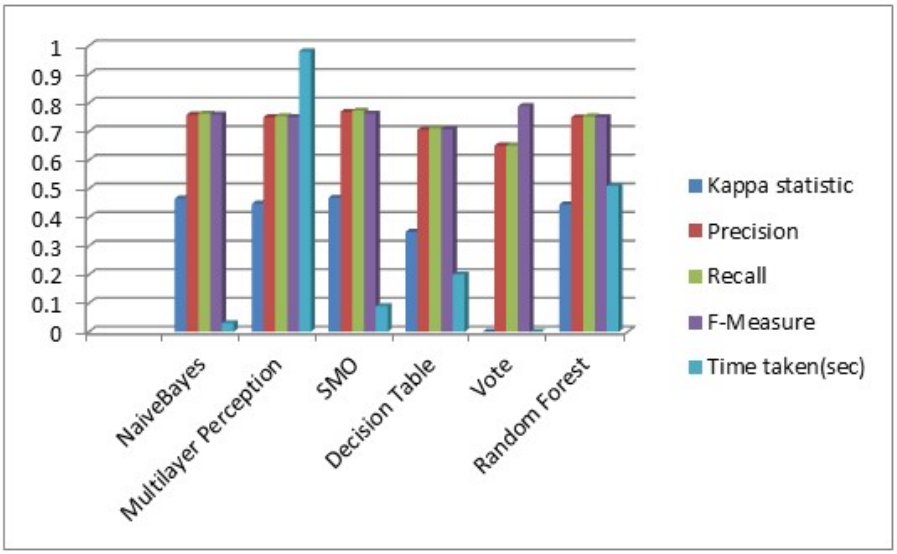



Figure 3. Bar graph of comparisons of diabetes diseases evaluation

*C. Result of Liver Diseases*

For liver disease classification analysis, dataset is retrieved from UCI repository. Total number of instances are 768 and number of features are 10. Data set contains both nominal and numerical data. The classification algorithm SMO performed best in terms of correctly classified instances. Its classification accuracy is 76.44%. The simulation results are shown in figure 4.

Table 7 provide us evaluation matrix performance of different algorithm on liver characteristic dataset. 10 Features were involved in liver dataset which are Age of the patient, Gender of the patient, TB Total Bilirubin, DB Direct Bilirubin, Alkaline Phosphotase, Sgpt Alanine Aminotransferase, Sgot Aspartate Aminotransferase, TP Total Protiens, ALB Albumin, A/G Ratio Albumin and Globulin Ratio. On this feature naïvebayes and sequential minimal optimization give better result. From these two SMO give slightly better results.

Table 7. Accuracy Evaluation of Data Mining and Machine Learning Algorithm on Liver Disease

| Algorithm | Correctly Classified Instances | Incorrectly Classified Instances | Kappa statistic | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|
| NaiveBayes | 586 | 182 | 0.4664 | 0.759 | 0.763 | 0.760 |
| Multilayer Perception | 579 | 189 | 0.4484 | 0.750 | 0.754 | 0.751 |
| SMO | 594 | 174 | 0.4682 | 0.769 | 0.773 | 0.763 |
| Decision Table | 547 | 221 | 0.3492 | 0.706 | 0.712 | 0.708 |
| Vote | 500 | 268 | 0 | 0.651 | 0.651 | 0.789 |
| Random Forest | 578 | 190 | 0.4459 | 0.749 | 0.753 | 0.750 |
| J48 | 567 | 201 | 0.4146 | 0.735 | 0.738 | 0.736 |

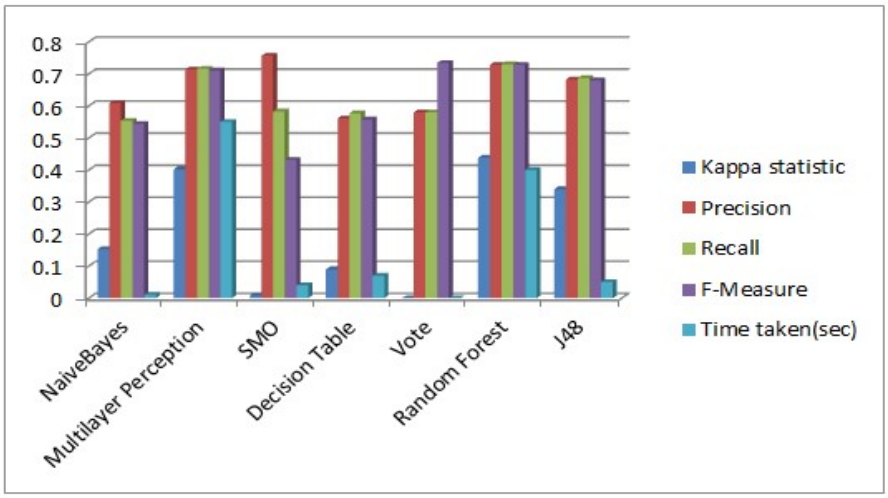



Figure 4. Bar graph of comparisons of liver diseases evaluation

*D. Result of Kidney Diseases*

Graphs are used for visual representation of the simulation results which are shown in figure 5. UCI repository is used for retrieving CKD data set. The data set contains 400 numbers of instances while the total number of features are 25.

Characteristics of Kidney datasets are age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edem, anemia.

Classification algorithms are applied and result shows that Random Forest classifier performed best for CKD prediction. Results obtained after simulation are summarized in the figure 5. Table 8 will give you detail insight evaluation model values on the dataset of kidney disease.

Table 8. Accuracy Evaluation of Data Mining and Machine Learning Algorithm on Kidney Dataset

| Algorithm | Correctly Classified Instances | Incorrectly Classified Instances | Kappa statistic | Precision | Recall | F-Measure |
|---|---|---|---|---|---|---|
| NaiveBayes | 380 | 20 | 0.8961 | 0.956 | 0.950 | .951 |
| Multilayer Perception | 399 | 1 | 0.9947 | 0.998 | 0.988 | 0.988 |
| SMO | 391 | 9 | 0.9526 | 0.979 | 0.978 | 0.978 |
| Decision Table | 396 | 4 | 0.9786 | 0.990 | 0.990 | 0.990 |
| Vote | 250 | 150 | 0 | 0.625 | 1.000 | 0.769 |
| Random Forest | 400 | 0 | 1 | 1.000 | 1.000 | 1.000 |
| J48 | 396 | 4 | 0.9786 | 0.990 | 0.990 | 0.990 |

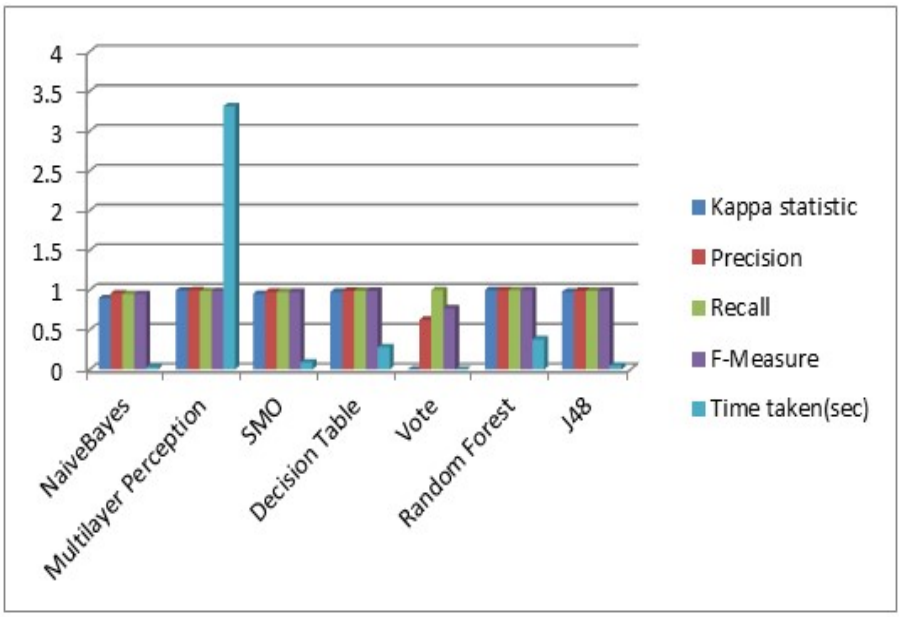



Figure 5. Bar graph of comparisons of kidney diseases evaluation

On the characteristics of Kidney disease we clearly see Random Forest giving us best precision and recall. Overall F measure of Random forest outperform other algorithm. Decision table provides us good accuracy but after Random forest.

## 3. Conclusion

Several data mining algorithms perform well for predicting various diseases i.e. heart, liver, kidney and diabetes. It is analyzed from existing literature that SVM and Naive bayes are most commonly and widely used algorithms for disease prediction. Accuracy of both algorithms outperforms as compared to other algorithms. KNN, SMO and Random Forest algorithms are also used but due to their complexity they are not widely accepted and preferred for disease prediction. Statistical models are failed to deal with big data. A significant role is played by data mining for dealing with huge data sets. In this paper, first different prediction techniques for

different diseases are reviewed. Further, data mining algorithms Naive Bayes, J48, REFTree, SMO, Multilayer Perceptron, and Vote are implemented in Weka using different data sets obtained from UCI respiratory. Parameters used for performance analysis of algorithms are correctly classified instances, precision, recall and F-Measure. After simulation results and discussion it is concluded that Random Forest algorithm shows best accuracy for heart, liver and kidney disease prediction. SMO performs best for diabetes prediction with an accuracy of 77.34%. In future work we could apply big data using spark [63-69] and deep learning approaches inspired by recent work [70-75].

## 4. References


[1] Adhikary, Junas, J. Han, and K. Koperski. "Knowledge Discovery in Spatial Databases-Progress and Challenges." *School of Computing Science, Simon Fraser University (1996)*.

[2] Kunwar, Veenita, et al. "Chronic Kidney Disease analysis using data mining classification techniques." *2016 6th International ConferenceCloud System and Big Data Engineering (Confluence). IEEE*, 2016.

[3] Durairaj, M., Ranjani, V. (2013). Data mining applications in healthcare sector: a study. *International journal of scientific technology research*, 2(10), 29-35.

[4] Tomar, D., Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, 5(5), 241-266.

[5] R Ricca F, Tonella P, Girardi C, et al. An empirical study on keyword-based web site clustering. Program Comprehension, 2004. *Proceedings. 12th IEEE International Workshop on. IEEE*; 2004.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982]*.

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," *in 2016 3rd International Conference on Electrical Engineering and Information Communication Technology(ICEEICT). IEEE, 2016, pp. 1–5*

[9] G. Shanmugasundaram, V. M. Selvam, R. Saravanan, and S. Balaji, "An investigation of heart disease pre-diction techniques," *in 2018 IEEE International Conference on System, Computation, Automation and Net-working (ICSCA). IEEE, 2018, pp. 1–6.*

[10] M. Fatima and M. Pasha, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications, vol. 9, no. 01, p. 1, 2017.*

[11] W.-L. Zuo, Z.-Y. Wang, T. Liu, and H.-L. Chen, "Effective detection of parkinson's disease using an adaptive fuzzy k-nearest neighbor approach," *Biomedical Signal Processing and Control, vol. 8, no. 4, pp. 364–373, 2013.*

[12] T. Revathi and S. Jeevitha, "Comparative study on heart disease prediction system using data mining tech- niques, "*International Journal of Science and Research (IJSR) ISSN (Online), pp. 2319–7064, 2013.*

[13] A. Taneja et al. "Heart disease prediction system using data mining techniques," *Oriental Journal of Computer science and technology, vol. 6, no. 4, pp. 457–466, 2013.*

[14] N. Masih and S. Ahuja, "Prediction of heart diseases using data mining techniques: Application on framingham heart study," *International Journal of Big Data and Analytics in Healthcare (IJBDAH), vol. 3, no. 2, pp. 1–9, 2018*

[15] V. Krishnaiah, D. G. Narsimha, and D. N. S. Chandra, "Diagnosis of lung cancer prediction system using data mining classification techniques," *International Journal of Computer Science and Information Technologies, vol. 4, no. 1, pp. 39–45, 2013.*

[16] P. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study," *ARPN Journal of Engineering and Applied Science, vol. 10, no. 1, pp. 8–13, 2015.*

[17] P. Agrawal and A. Dewangan, "A brief survey on the techniques used for the diagnosis of diabetes-mellitus," *Int. Res. J. of Eng. and Tech. IRJET, vol. 2, pp. 1039–1043, 2015.*

[18] S. Vijayarani, S. Dhayanand, and M. Phil, "Kidney disease prediction using svm and ann algorithms," *International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2015.*

[19] B. Boukenze, H. Mousannif, and A. Haqiq, "Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease," *Int. Journal of Database Managment systems, vol. 8, no. 30, pp. 1–9, 2016.*

[20] L. Jena and N. K. Kamila, "Distributed data mining classification algorithms for prediction of chronic-kidney- disease," *Int. J. Emerg. Res. Manag. Technology, vol. 9359, no. 11, pp. 110–118, 2015.*

[21] S. Vijayarani and S. Dhayanand, "Liver disease prediction using svm and na ıve bayes algorithms," *International Journal of Science, Engineering and Technology Research (IJSETR), vol. 4, no. 4, pp. 816–820, 2015.*

[22] P. M. Patil, "Review on prediction of chronic kidney disease using data mining techniques," *International Journal of Computer Science and Mobile Computing, vol. 5, no. 5, p. 135, 2016.*

[23] A. Aneeshkumar and C. J. Venkateswaran, "A novel approach for liver disorder classification using data mining techniques," *Engineering and Scientific International Journal, vol. 2, no. 1, pp. 15–18, 2015*

[24] D. Sindhuja and R. J. Priyadarsini, "A survey on classification techniques in data mining for analyzing liver disease disorder," *International Journal of Computer Science and Mobile Computing, vol. 5, no. 5, pp. 483–488, 2016.*

[25] M. Abdar, S. R. N. Kalhori, T. Sutikno, I. M. I. Subroto, and G. Arji, "Comparing performance of data mining algorithms in prediction heart diseases." *International Journal of Electrical Computer Engineering (2088-8708), vol. 5, no. 6, 2015.*

[26] J. Thomas and R. T. Princy, "Human heart disease prediction system using data mining techniques," *in 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT). IEEE, 2016, pp. 1–5*

[27] I. A. Zriqat, A. M. Altamimi, and M. Azzeh, "A comparative study for predicting heart diseases using data mining classification methods," arXiv preprint arXiv:1704.02799, 2017

[28] M. Rathi and B. Narasimhan, "Data mining, soft computing, machine learning and bio-inspired computing for heart disease classification/prediction–a review," *International Journal Advanced Research in Computer Science and Software Engineering, vol. 7, no. 4, 2017.*

[29] I. E. Emre, N. Erol, Y. I. Ayhan, Y. Ozkan, and C. Erol, "The analysis of the effects of acute rheumatic fever in childhood on cardiac disease with data mining," *International journal of medical informatics, vol. 123, pp. 68–75, 2019.*

[30] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," Procedia Computer Science, vol. 82, pp. 115–121, 2016.

[31] T. Daghistani and R. Alshammari, "Diagnosis of diabetes by applying data mining classification techniques," *International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 7, pp. 329–332, 2016*

[32] A. Azrar, M. Awais, Y. Ali, and K. Zaheer, "Data mining models comparison for diabetes prediction," *International Journal of Advanced Computer Science and Applications, vol. 9, no. 8, pp. 320–323, 2018.*

[33] R. C. S. Mary and B. S. Kumar, "Comparison of various data mining algorithms in the prediction of risk for gestational diabetes," *International Journal of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 8, p. 74, 2018*

[34] S. S. Shinde, R. M. Rajmane, S. S. Chindage, S. S. Gundale, and U. B. Mane, "A survey on prediction of diabetes using data mining," 2018

[35] M. L. Z. Alkaragole and A. P. S. Kurnaz, "Comparison of data mining techniques for predicting diabetes or prediabetes by risk factors," 2019.

[36] M. Abdar, M. Zomorodi-Moghadam, R. Das, and I.-H. Ting, "Performance analysis of classification algorithms on early detection of liver disease," Expert Systems with Applications , vol. 67, pp. 239–251, 2017.

[37] M. Hassoon, M. S. Kouhi, M. Zomorodi-Moghadam, and M. Abdar, "Rule optimization of boosted c5. 0 clas- sification using genetic algorithm for liver disease prediction," in 2017 International Conference on Computer and Applications (ICCA) . *IEEE, 2017, pp. 299–305*

[38] M. Pasha and M. Fatima, "Comparative analysis of meta learning algorithms for liver disease detection." JSW , vol. 12, no. 12, pp. 923–933, 2017.

[39] S. Kumar and S. Katyal, "Effective analysis and diagnosis of liver disorder by data mining," *in 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, 2018, pp. 1047–1051.*

[40] I. Arshad, C. Dutta, T. Choudhury, and A. Thakral, "Liver disease detection due to excessive alcoholism using data mining techniques,"
in 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE). *IEEE, 2018, pp. 163–168*

[41] M. Mesgar and M. R. Shahraki, "Evaluation of data mining algorithms for detection of liver disease," *Journal of Payavard Salamat, vol. 13, no. 1, pp. 0–0, 2019.*

[42] A. N. Arbain and B. Y. P. Balakrishnan, "A comparison of data mining algorithms for liver disease prediction on imbalanced data," *International Journal of Data Science and Advanced Analytics, vol. 1, no. 1, pp. 1–11, 2019*

[43] C.-C. Wu, W.-C. Yeh, W.-D. Hsu, M. M. Islam, P. A. A. Nguyen, T. N. Poly, Y.-C. Wang, H.-C. Yang, and Y.-C. J. Li, "Prediction of fatty liver disease using machine learning algorithms," *Computer methods and programs in biomedicine, vol. 170, pp. 23–29, 2019.*

[44] S. Vijayarani and S. Dhayanand, "Data mining classification algorithms for kidney disease prediction," *International Journal on Cybernetics Informatics (IJCI) , vol. 4, no. 4, pp. 13–25, 2015.*

[45] H. Alasker, S. Alharkan, W. Alharkan, A. Zaki, and L. S. Riza, "Detection of kidney disease using various intelligent classifiers," *in 2017 3rd International Conference on Science in Information Technology (ICSITech). IEEE, 2017, pp. 681–684*

[46] M. S. Gharibdousti, K. Azimi, S. Hathikal, and D. H. Won, "Prediction of chronic kidney disease using data mining techniques," *in IIE Annual Conference. Proceedings. Institute of Industrial and Systems Engineers (IISE), 2017, pp. 2135–2140.*

[47] S. Zeynu and S. Patil, "Survey on prediction of chronic kidney disease using data mining classification tech- niques and feature selection," *International Journal of Pure and Applied Mathematics, vol. 118, no. 8, pp. 149–156, 2018.*

[48] E. Avci, S. Karakus, O. Ozmen, and D. Avci, "Performance comparison of some classifiers on chronic kidney disease data," *in 2018 6th International Symposium on Digital Forensic and Security (ISDFS). IEEE, 2018, pp. 1–4*

[49] K. A. Otunaiya and G. Muhammad, "Performance of datamining techniques in the prediction of chronic kidney disease," 2019.

[50] G. Murshid, T. Parvez, N. Fezal, L. Azaz, and M. Asif, "Data mining techniques to predict chronic kidney disease," 2019.

[51] Ali, U., Shamsi, M. H., Nabeel, M., Hoare, C., Alshehri, F., Mangina, E., & Odonnell, J. (2019, November). Comparative analysis of prediction algorithms for building energy usage prediction at an urban scale. *In Journal of Physics: Conference Series (Vol. 1343, No. 1, p. 012001). IOP Publishing.*

[52] Ather, S., Muslin-Ud-Din, H., Nabeel, M., Ahsan, M., & Hassan, B. (2019). Several Adaptive Replica Synchronization Approaches for Distributed file System. VAWKUM Transactions on Computer Sciences, 7(1), 1-8.

[53] Tariq, Z. B., Arshad, N., & Nabeel, M. (2015, May). Enhanced LZMA and BZIP2 for improved energy data compression. *In 2015 International Conference on Smart Cities and Green ICT Systems (SMARTGREENS) (pp. 1-8). IEEE.*

[54] Nabeel, M., Javed, F., & Arshad, N. (2013). Towards Smart Data Compression for Future Energy Management System. *In Fifth International Conference on Applied Energy (pp. 1-4).*

[55] Nabeel, M., & Hassan, M. T. Detection of Missing Values from Big Data of Self Adaptive Energy Systems.

[56] A. A. Rehma, M. J. Awan, and I. Butt, "Comparison and Evaluation of Information Retrieval Models," *VFAST Transactions on Software Engineering,* vol. 6, no. 1, pp. 7-14, 2018.

[57] Y. Ali, A. Farooq, T. M. Alam, M. S. Farooq, M. J. Awan, and T. I. Baig, "Detection of Schistosomiasis Factors Using Association Rule Mining," *IEEE Access,* vol. 7, pp. 186108-186114, 2019.

[58] T. M. Alam, and M. J. Awan, "Domain analysis of information extraction techniques," *International Journal of Multidisciplinary Sciences and Engineering,* vol. 9, pp. 1-9, 2018.

[59] M. Anam, V. a/p Ponnusamy, M. Hussain, M. Waqas Nadeem, M. Javed, H. Guan Goh, and S. Qadeer, "Osteoporosis Prediction for Trabecular Bone using Machine Learning A Review," *Computers, Materials & Continua,* vol. 67, no. 1, pp. 89-105, 2021.

[60] M. Gupta, R. Jain, S. Arora, A. Gupta, M. Javed Awan, G. Chaudhary, and H. Nobanee, "AI-enabled COVID-9 Outbreak Analysis and Prediction: Indian States vs. Union Territories," *Computers, Materials & Continua,* vol. 67, no. 1, pp. 933-950, 2021.

[61] R. Javed, T. Saba, S. Humdullah, N. S. M. Jamail, and M. J. Awan, "An Efficient Pattern Recognition Based Method for Drug-Drug Interaction Diagnosis." pp. 221-226.

[62] A. T. Nagi, M. J. Awan, R. Javed, and N. Ayesha, "A Comparison of Two-Stage Classifier Algorithm with Ensemble Techniques On Detection of Diabetic Retinopathy." pp. 212-215.

[63] M. Javed Awan, M. Shafry Mohd Rahim, H. Nobanee, A. Yasin, O. Ibrahim Khalaf, and U. Ishfaq, "A Big Data Approach to Black Friday Sales," *Intelligent Automation & Soft Computing,* vol. 27, no. 3, pp. 785-797, 2021.

[64] . M. Javed Awan, M. Shafry Mohd Rahim, H. Nobanee, A. Munawar, A. Yasin, and A. Mohd Zain Azlanmz, "Social Media and Stock Market Prediction: A Big Data Approach," *Computers, Materials & Continua,* vol. 67, no. 2, pp. 2569-2583, 2021

[65] H. M. Ahmed, M. J. Awan, N. S. Khan, A. Yasin, and H. M. F. Shehzad, "Sentiment Analysis of Online Food Reviews using Big Data Analytics," *Elementary Education Online,* vol. 20, no. 2, pp. 827-836, 2021.

[66] M. J. Awan, R. A. Khan, H. Nobanee, A. Yasin, S. M. Anwar, U. Naseem, and V. P. Singh, "A Recommendation Engine for Predicting Movie Ratings Using a Big Data Approach," *Electronics,* vol. 10, no. 10, pp. 1215, 2021.

[67] Aftab, M. O., Awan, M. J., Khalid, S., Javed, R., & Shabir, H. (2021, April). Executing Spark BigDL for Leukemia Detection from Microscopic Images using Transfer Learning. *In 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA) (pp. 216-220). IEEE.*

[68] M. J. Awan, M. A. Khan, Z. K. Ansari, A. Yasin, and H. M. F. Shehzad, "Fake Profile Recognition using Big Data Analytics in Social Media Platforms," *Interational Journal of Computer Applications in Technology,* vol. press, 2021.

[69] A. Khalil, M. J. Awan, A. Yasin, V. P. Singh, and H. M. F. Shehzad, "Flight Web Searches Analytics through Big Data," *Interational Journal of Computer Applications in Technology*, (in press).

[70] M. Awan, M. Rahim, N. Salim, A. Ismail, and H. Shabbir, "Acceleration of knee MRI cancellous bone classification on google colaboratory using convolutional neural network," *Int. J. Adv. Trends Comput. Sci,* vol. 8, pp. 83-88, 2019.

[71] M. J. Awan, M. S. M. Rahim, N. Salim, M. A. Mohammed, B. Garcia-Zapirain, and K. H. Abdulkareem, "Efficient Detection of Knee Anterior Cruciate Ligament from Magnetic Resonance Imaging Using Deep Learning Approach," *Diagnostics (Basel),* vol. 11, no. 1, Jan 11, 2021.

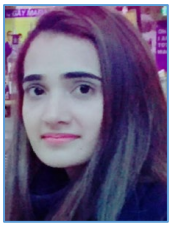
[72] Abdullah, Y. Awais, M. J. Awan, M. F. Shehzad, and M. Ashraf, "Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory," *International Journal of Emerging Technologies in Learning* vol. 11, no. 2, pp. 209-212, 2020.

[73] M. J. Awan, A. Raza, A. Yasin, H. M. F. Shehzad, and I. Butt, "The Customized Convolutional Neural Network of Face Emotion Expression Classification," *Annals of the Romanian Society for Cell Biology,* vol. 25, no. 6, pp. 5296-5304, 2021.

[74] A. Mujahid, M. J. Awan, A. Yasin, M. A. Mohammed, R. Damaševičius, R. Maskeliūnas, and K. H. Abdulkareem, "Real-Time Hand Gesture Recognition Based on Deep Learning YOLOv3 Model," *Applied Sciences,* vol. 11, no. 9, 2021.

[75] R. Mubashar, M. Javed Awan, M. Ahsan, A. Yasin, and V. Partap singh, "Efficient Residential Load Forecasting using Deep Learning Approach," *International Journal of Computer Applications in Technology,* vol. (in press), 2021.

**Muhammad Nabeel** is currently working as an Assistant Professor in Software Engineering Department, University of Management and Technology (UMT) Lahore Pakistan. My area of interests is Data Science, deep learning, machine learning, and Intelligent Computing.



**Shumaila Majeed** is currently a M. Phil student at the Department of Computer Science, Fatima Jinnah University, Rawalpindi. She obtained her Bachelor's degree in Computer Science at International Islamic University, Islamabad in 2015. Her area of interests are deep learning, machine learning, and computer vision. Her current research focus is in the area of transfer learning methods in biomedical imaging. Currently, she is working on Brain Tumour detection and classification using transfer learning.

**Mazhar Javed Awan** is an Assistant Professor of Software Engineering Department at the University of Management & Technology (UMT) Lahore, Pakistan. He has overall 18 years of teaching experience in various Institutes. His Ph.D degree is from University Teknologi Malyasia (UTM). His areas of research interest include Data sciences, Big Data Analytics, Deep learning in medical images, Natural language processing and Machine learning. He is reviewer of many WOS and Scopus Indexed journals like IEEE Internet of things and CMC Journals. Besides research he is also a keynote speaker at National and International level at various conferences and workshops related to Data Science and Big Data.

**Hooria Muslih-ud-Din** has a bachelor's degree with Computer Science from FAST-National University, Lahore, Pakistan. She is currently working in Atheneum-Partners as a Survey Programmer specifically working in Python Development. She has six months' work experience at Learning Hub Pvt Ltd and completed an internship at Falconic Tech in Flutter.

**Mashal Wasique** is currently a M. Phil student at Department of Computer Science, Fatima Jinnah University , Rawalpindi. She obtained her Bachelor's degree in Computer Science at University of Gujrat, Gujrat in 2017. Her area of interests is adhoc networks, security in adhoc networks, machine learning, computer vision. Her current research focus is in the area of secure routing in MANET. Currently, she is working on attack detection and mitigation in mobile adhoc networks.

**Rabia Nasir** is currently working in the Anesthesia Department of Lahore General Hospital in Lahore Pakistan. She completed her Bachelor of Medicine and a Bachelor of Surgery (M.B.B.S) from Services Institute of Medical Sciences (SIMS), Pakistan.

.