International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Diagnosis of Chronic Kidney Disease Using Machine Learning Methods

Rasheed Shaikh[1], Kurhe P.V.[2], Dr. Umesh Pawar[3] , Ramesh Daund[4]

[1]Department of Computer Engineering, SND COE& RC, Yeola, Nashik
[2]Departmetn of Computer Engineering, SND COE & RC, Yeola, Nashik
[3]Department of Computer Engineering, SND COE& RC, Yeola, Nashik
[4]Department of Computer Engineering, SND COE& RC, Yeola, Nashik

---

Abstract

As a leading cause of morbidity and mortality, chronic kidney disease (CKD) is a concern for global health. Effective patient management of CKD depends on an accurate and quickly determined diagnosis. Machine learning (ML) techniques have shown promise in a number of medical domains and may improve the diagnosis of CKD. A unique ML model for the diagnosis of CKD is presented in this paper. A variety of machine learning (ML) algorithms, such as decision trees, support vector machines, and artificial neural networks, are used in the suggested model. It is especially made to examine clinical and laboratory information that is frequently accessible for the diagnosis of CKD. For accurate diagnosis, the most informative variables are found using feature selection techniques. An extensive CKD dataset made up of patient demographics, medical history, laboratory test results, and imaging data is used for training and validation in order to create and evaluate the model's performance. The model's diagnostic abilities are assessed using performance evaluation criteria like accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve. Initial results show good performance, displaying excellent sensitivity and accuracy in CKD diagnosis. The proposed approach may help medical practitioners diagnose patients quickly and accurately, enable effective therapy actions, and lower the risk of illness progression.

Key Words: Machine Learning, Medical Diagnosis, CKD Diagnosis, Medical Health.

1. INTRODUCTION

The term "chronic kidney disease" (CKD) describes the kidneys' inability to carry out its typical blood filtration function and other tasks [1]. The steady decrease of kidney cells over time is referred to as "chronic" in medical terminology. In this case of severe renal failure, the body has a considerable fluid buildup and the kidneys are no longer able to filter blood. The body's levels of potassium and calcium salts increase drastically as a result. The body experiences a number of extra issues when these salts are present in large amounts. The kidneys' main job is to filter extra water and waste from the blood [3]. The minerals and salts in our bodies must be balanced for this process to be effective. In order to regulate blood pressure, activate hormones, and produce red blood cells, among other things, the appropriate salt balance is necessary [18]. The body experiences a number of extra issues when these salts are present in large amounts. The kidneys' main job is to filter extra water and waste from the blood [25]. The minerals and salts in our bodies must be balanced for this process to be effective. In order to regulate blood pressure, activate hormones, and produce red blood cells, among other things, the appropriate salt balance is necessary. Women who have a high calcium intake also tend to have cystic ovaries. A sudden illness or a medicine allergy can both be brought on by CKD. The medical term for this illness is acute kidney injury (AKI). Heart problems and heart attacks can be brought on by high blood pressure.

Particularly in remote and hard-to-reach areas, developing nations struggle with a lack of primary care professionals or unstable primary healthcare [24]. Early detection of chronic kidney disease may help primary care doctors contact populations that are located in remote and difficult-to-reach areas, especially in light of the prevalence of mobile technology and Internet connectivity today. Additionally, software-assisted CKD diagnoses may boost clinical evaluations' trust, which would help solve the issue of poor primary healthcare in underdeveloped nations. Software tools have been created to help doctors monitor and diagnose CKD [4]. Techniques linked to machine learning have also been used to improve the ability of software systems to track and identify chronic illnesses [5]. Due to the unique issues these populations experience, such as insufficient primary healthcare, the current study focuses on an appropriateness analysis of machine learning approaches and discusses how software systems might be used to aid in CKD's early identification in developing nations [8]. In the world of medicine, machine learning and data mining have proven to be quite beneficial.

Medical sciences as a result depend heavily on techniques and well-organized data for the analysis, prediction, and diagnosis of diseases [10]. They extract patterns from the data, and the patients can employ these patterns to survive. Numerous categorization models have been successfully used and deployed for many different reasons of Among them are Support Vector Machine (SVM), Logistic Regression (LG), and Naive Bayes (NB) [12]. A sort of ensemble machine learning technique called boosting algorithms turns a weak classifier into a strong model to obtain higher accuracy. In our

research, we will examine ensemble learning models that are based on trees and aim to outperform current models in terms of speed and accuracy. The categorization of machine learning is shown in figure 1 [4].

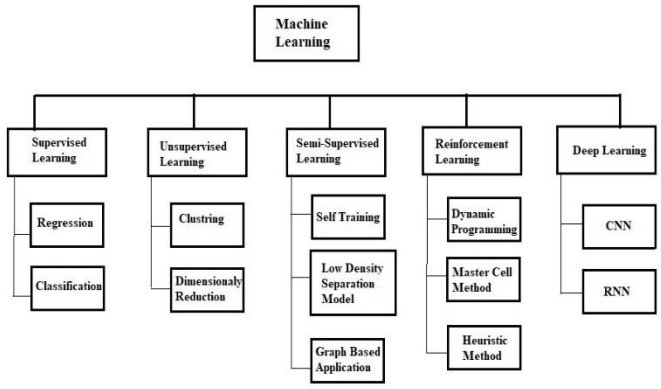

Figure 1: Machine Learning Categorization

2. PROBLEM STATEMENT

Primary care physicians are in short supply or the quality of basic healthcare is poor in developing countries, particularly in remote and inaccessible areas. Given the current prevalence of mobile technology and Internet access, primary care physicians may find it easier to use computer-assisted CKD early diagnosis to reach segments of the population living in remote and hard-to-reach areas. In addition, software-assisted CKD diagnoses may increase confidence in clinical assessments, which may help address the poor primary health care problem in developing countries.

3. PROPOSED MODEL

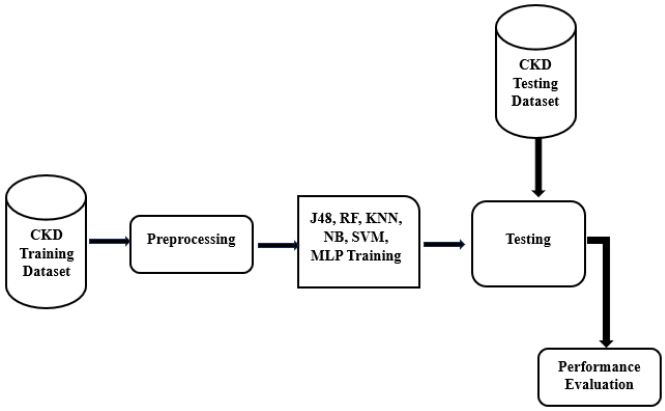

Figure 2: Proposed Model

Preprocessing, which involves preparing the raw data for machine learning algorithms, is an important stage in CKD prediction [9]. In addition to identifying pertinent features, scaling numerical variables, encoding categorical variables, correcting class imbalance, and dividing the dataset for training and testing, it also entails handling missing values, outliers, and inconsistencies. Preprocessing transforms the data into a useful format, improving the CKD prediction model's accuracy and dependability [16].

4. PROBLEM SOLVING APPROACH

4.1 J48 Algorithm: A popular machine learning method for categorical and continuous data processing is the J48 algorithm [12]. It is based on the C4.5 algorithm and has a wide range of applications. For instance, it is frequently used to classify E-governance data, evaluate clinical data for the diagnosis of coronary heart disease, and perform other related activities.

4.2 RF Algorithm: Popular supervised learning algorithm Random Forest is frequently used in machine learning. Both classification and regression issues can be solved with it [25]. The approach makes use of ensemble learning, a technique that combines different classifiers to tackle challenging issues and enhance model performance. In the case of Random Forest, it is made up of several decision trees that were constructed using various dataset subsets. The technique improves the model's forecast accuracy by averaging the results of these trees.

4.3 The K-Nearest Neighbor Algorithm: Non-parametric techniques like the K-nearest neighbour (K-NN) algorithm are frequently employed for classification and regression tasks [28]. The K nearest training instances in the feature space are taken into account while it functions. An object's class membership in a K-NN classification is decided by a majority vote from its K nearest neighbours. The object is simply assigned to the class of its one nearest neighbor if K is set to 1 [27].

4.4 Naïve Bayes Algorithm: The Naive Bayes algorithm uses the Bayes' Theorem to classify data and assumes predictor independence. It makes the assumption that a certain feature's presence in a class is unconnected to the existence of any other feature [32]. Based on the existence of various characteristics, this method determines the possibility that an object belongs to a particular class and then chooses the class with the highest probability.

4.5 SVM Algorithm: For classification and regression analysis, supervised learning models called support vector machines (SVMs) are used. By examining a set of training examples and categorizing new examples according to their attributes, SVMs create a model [6]. The algorithm attempts to show the differences between the various categories by modelling the samples as points in space. Then, based on their placement in relation to the gap, new examples are projected to fit into one of the categories by being mapped into this space.

4.6 MLP Algorithm: An artificial neural network with complete connectivity and feedforward is known as a multilayer perceptron (MLP). It is made up of several layers of perceptron's, each of which is coupled to every perceptron in the layer below it [8]. The term "vanilla" neural network is frequently used to describe MLPs, particularly when there is only one hidden layer. They are frequently employed for many different tasks, like as classification and regression, and they are

able to recognize intricate patterns and connections in the data. This algorithm's effectiveness is evaluated at that point.

5. PERFORMANCE METRICS

There are four performance evaluation indexes of the performance of English subordinate clause connective correction: accuracy rate (Accuracy), precision rate P (Precision), recalls rate R (Recall) and F value (F1) [12], and their formulas are as follows:

Accuracy = (TP + TN) / (TP + FN + FP + TN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1 = 2 * (Precision * Recall) / (Precision + Recall)

Here, TP (True Positive) represents the number of positive samples sorted correctly by the classifier, FN (False Negative) is the number of positive samples misclassified by the classifier, FP (False Positive) indicates the number of negative samples misclassified by the classifier, and TN (True Negative) is the number of negative samples sorted correctly by the classifier [13].

5.1 Accuracy: Accuracy is a measure of how well a classification model correctly predicts the overall instances. It calculates the ratio of correctly classified instances to the total number of instances in the dataset [34]. Accuracy is useful when the classes in the dataset are balanced. However, it can be misleading in the presence of class imbalance.

5.2 Precision: Precision is a measure of the model's ability to correctly identify positive instances from the total instances it predicts as positive. It is calculated as the ratio of true positive instances to the sum of true positive and false positive instances [27]. Precision is particularly useful when the focus is on minimizing false positives or when the cost of misclassification is high.

5.3 Recall: Recall, also known as sensitivity or true positive rate, measures the model's ability to identify positive instances correctly out of the total actual positive instances [4]. It is calculated as the ratio of true positive instances to the sum of true positive and false negative instances. Recall is important when the goal is to minimize false negatives or when the cost of missing positive instances is high.

5.4 F1 Score: The F1 score is a metric that combines both precision and recall into a single value. It provides a balance between precision and recall, giving equal importance to both metrics [11]. The F1 score is the harmonic mean of precision and recall and is calculated as 2 * (precision * recall) / (precision + recall). It is particularly useful when there is an uneven class distribution or when both false positives and false negatives need to be minimized.

6. DATASETS

Chronic Kidney Disease dataset is used in this study. It is extracted from Kaggle machine learning repository [10]. The data was taken over a 2-month period in India with 25 features (eg, red blood cell count, white blood cell count, etc). The target is the ‘classification’, which is either ‘ckd’ or ‘notckd’ - ckd=chronic kidney disease. There are 400 rows.

7. RESULTS AND DISCUSSIONS

7.1 Experiments Setup

We used the Weka tool [30] to evaluate our machine learning models, and a computer was used for the trials., which has the following specifications: Intel Core i5 (sixth generation or newer) or equivalent @ 3.60GHz 3.80 GHz, 16 GB, Windows 11 Home, 64-bit Operating System and x64-based processor. By using 10-fold cross-validation to gauge the models' effectiveness in the balanced dataset of 500 instances following SMOTE, the experimental findings were obtained. Table 1 shows that the ideal values for the ML model parameters that we experimented with.

Table -1: Performance of ML models using SMOTE and 10-Fold Cross-Validation

Models	Parameters
Naive Bayes	useKernelEstimator: False useSupervisedDiscretization: True
Support Vector Machine	eps = 0.001 gamma = 0.0 kernel type: linear loss = 0.1
K-Nearest Neighbour	k = 1 Search Algorithm: LinearNNSearch with Euclidean
J48	reducedErrorPruning: False saveInstanceData: False subtreeRaising: True
Random Forest	maxDepth = 0 numIterations = 100 numFeatures = 0

The PCA feature transformation, which creates rotational matrices with few correlations and are distinguished by a lower cumulative proportion of matrix diversity, is related to the Rotation Forest method's distinctive accuracy. This makes it easier for various, mutually independent DTs to emerge inside a Rotation Forest ensemble, which enhances its accuracy. The Rotation Forest produces more precise individual classifiers than AdaBoostM1 and Random Forest, as shown by the results.

Applied RF, K-Nearest Neighbour and Decision Tree, achieving an accuracy of 97%, 71.25% and 96.25%, respectively. Our proposed models performed an accuracy of 97.4%, 98.4% and 97.4% for the LR, K-NN and Decision Tree models, respectively. We can observe that our proposed models demonstrate slightly better accuracy rates than the comparable research works, except for our k-NN model, which outperforms

with a performance gap of 26.15% concerning the respective model of the research work. The comparison of different ML models shows mentioned in Table 2.

Table -1: Performance of ML models using SMOTE and 10-Fold Cross-Validation

Algorithm	Accuracy	Precision	Recall	F1 Score
J48	0.99	0.98	0.97	0.98
RF	0.97	0.99	0.98	0.97
NB	0.80	0.84	0.80	0.79
SVM	0.90	0.92	0.90	0.91
MLP	0.85	0.89	0.85	0.84
KNN	0.85	0.89	0.85	0.84

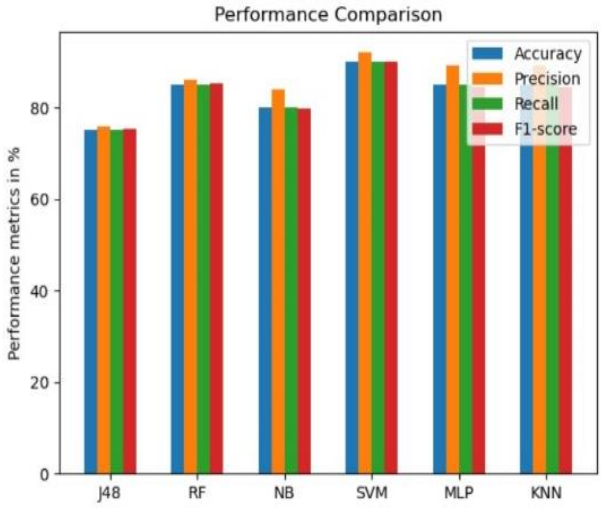

Figure 3 Performance Metrics Chart

3. CONCLUSIONS

Chronic kidney disease prediction is critical because it is now the main cause of death. In this paper we present a deep literature survey on system for diagnosis of CKD. Depending on the number of parameters evaluated during the classification, machine learning approaches provide varying levels of accuracy for CKD diagnosis. In future the machine learning algorithms can be implemented in such a way to improve the accuracy. In the context of developing countries, the costs resulting from the usage of software to assist in CKD diagnoses needs to be as low as possible, especially in hard-to-reach and rural settings. The number of CKD attributes used during CKD risk classifications impacts the cost of usage and the performance of the classifiers. The machine learning techniques present different levels of accuracy for the CKD diagnosis depending on the number of attributes considered during the classification. In this study, the J48 decision tree and RF exhibited the best performance using the CKD dataset, comprising of hypertension, DM, creatinine, urea, albuminuria, age, gender, and GFR attributes. These attributes are commonly used by nephrologists to diagnose CKD in developing countries. Nevertheless, the RF machine learning technique usually conducts more complex evaluations, making the interpretation of the classification results by physicians difficult. The application of algorithms (interpreters) is required to interpret the results before presenting them to primary care physicians. Conversely, the J48 decision tree addresses the interpretation problem still presenting nearly perfect agreement with experienced nephrologists who has treated CKD patients in developing countries for more than 30 years. In addition, critical misleading classifications are not presented by the J48 decision tree classifier when evaluating the CKD dataset subjects.

ACKNOWLEDGEMENT

The heading should be treated as a 3rd level heading and should not be assigned a number.

REFERENCES

[1] Mirza Muntasir Nishat et al 2021 “A Comprehensive Analysis on Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms”.13 August 2021

[2] BILAL KHAN et al 2020 “An Empirical Evaluation of Machine Learning Techniques for Chronic Kidney Disease Prophecy” March 18,2020.

[3] Adeeba Azmi1, Amiksha Hingu2, Ruchi Dholaria3, Ms. Alvina Alphonso. 2020 “Chronic Kidney Disease Prediction using Data Mining and Machine Learning” March 2020.

[4] Arif-Ul-Islam, Shamim H Ripon, 2019“Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree” February 2019.

[5] S.Belina V.J Sara, Dr.K.Kalaiselvi, “Ant Colony Optimization (ACO) Based Feature Selection And Extreme Learning Machine (ELM) For Chronic Kidney Disease Detection”. 2018.

[6] Doni Aprilianto 2020, “SVM Optimization with Correlation Feature Selection Based Binary Particle Swarm Optimization for Diagnosis of Chronic Kidney Disease”, September 2020

[7] Shanila Yunus Yashfi et al 2020 "Risk Prediction Of Chronic Kidney Disease Using Machine Learning Algorithms". IEEE, October 2020

[8] Jing Xiao et al 2019,"Comparison and development of machine learning tools in the prediction of chronic kidney disease progression", BMC 2019

[9] U Abinayan et al 2021 "Noval Approach For Chronic Kidney Disease Using Machine Learning Methodology",Journal of Physics: Conference Series, Volume 1916, 2021 International Conference on Computing, Communication, Electrical and Biomedical Systems (ICCCEBS) 2021 March 2021, Coimbatore, India

[10] I.A. Pasadana et al 2018, "Chronic Kidney Disease Prediction by Using Different Decision Tree Techniques" The International Conferenczon Computer Science and Applied Mathematic October 2018

[11] Qin, Jiongming; Chen, Lin; Liu, Yuhua; Liu, Chuanjun; Feng, Changhao; Chen, Bin 2019." A Machine Learning Methodology for Diagnosing Chronic Kidney Disease." IEEE 2019

[12] Nusrat Tazin, Shahed Anzarus Sabab, Muhammed Tawfiq Chowdhury Diagnosis of Chronic Kidney Disease Using Effective Classification and Feature Selection Technique, 2021

[13] eGFR Calculator. https: / / www.kidney.org / apps / professionals/egfr-calculator

[14] Kidney disease assistant App. https : / /appadvice.com / app/kidney-disease-assistant/830127960

[15] D. W. Cockcroft, and M. H. Gault, "Prediction of creatinine clearance from serum creatinine," Nephron, vol. 16, no. 1, pp. 31–41, 1976, 10.1159/000180580.

[16] A. S. Levey, J. P. Bosh, B. J. Lewis, T. Greene, N. Rogers, and D. Roth, "A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation," Ann Intern Med, vol. 130, no. 6, pp. 461–470, March 1999, 10.7326/0003-4819-130-6-199903160-00002.

[17] A. S. Levey, C. Schmid, Y. Zhang, A. R. Castro, H. Feldman, J. Kusek, P. Eggers, F. Van Lente, T. Greene, and J. Coresh, "A new equation to estimate glomerular filtration rate," Ann Intern Med, vol. 150, no. 9, pp. 604–612, May 2009, 10.7326/0003-4819-150-9-200905050-00006.

[18] G. J. Schwartz, A. Muñoz, M. F. Schneider, R. H. Mak, F. Kaskel, B. A. Warady, and S. L. Furth, "New equations to estimate gfr in children with ckd," Journal of the American Society of Nephrology, vol. 20, no. 3, pp. 629–637, March 2009, 10.7326/0003-4819-150-9-200905050-00006.

[19] A. Sobrinho, L. D. Silva, A. Perkusich, M. E. Pinheiro, and P. Cunha, "Design and evaluation of a mobile application to assist the self- monitoring of the chronic kidney disease in developing countries," BMC Medical Informatics and Decision Making, vol. 18, no. 1, pp. 1–14, Jan. 2018, 10.1186/s12911-018-0587-9.

[20] L. Liu, Y. Ni, N. Zhang, and J. Pratap, "Mining patient-specific and contextual data with machine learning technologies to predict cancellation of children's surgery," International Journal of Medical Informatics, vol. 129, pp. 234–241, June 2019, 10.1016/j.ijmedinf.2019.06.007.

[21] K. Topuz, F. D. Zengul, A. Dag, A. Almehmi, and M. B. Yildirim, "Predicting graft survival among kidney transplant recipients: A bayesian decision support model," Decision Support Systems, vol. 106, pp. 97–109, Feb. 2018, 10.1016/j.dss.2017.12.004.

[22] F. F. Jahantigh, B. Malmir, and B. A. Avilaq, "A computer-aided diagnostic system for kidney disease," Kidney Research and Clinical Practice, vol. 36, no. 1, pp. 29–38, March 2017, 10.23876/j.krcp.2017.36.1.29.

[23] J. Neves, M. R. Martins, J. Vilhena, J. Neves, S. Gomes, A. Abelha, J. Machado, and H. Vicente, "A soft computing approach to kidney diseases evaluation," Journal of Medical Systems, vol. 39, no. 10, pp. 131, Aug. 2015, 10.1007/s10916-015-0313-4.

[24] H. Polat, H. D. Mehr, and A. Cetin, "Diagnosis of chronic kidney disease based on support vector machine by feature selection methods," Journal of Medical Systems, vol. 41, no. 55, pp. 1–11, Feb. 2017, 10.1007/s10916- 015-0313-4.

[25] Arif-Ul-Islam and S. H. Ripon, "Rule Induction and Prediction of Chronic Kidney Disease Using Boosting Classifiers, Ant-Miner and J48 Decision Tree," 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 2019, pp. 1-6

[26] J. Aljaaf et al., "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics," 2018 IEEE Congress on Evolutionary Computation (CEC), Rio de Janeiro, 2018, pp. 1-9.

[27] V. Ravindra, N. Sriraam and M. Geetha, "Discovery of significant parameters in kidney dialysis data sets by K-means algorithm," International Conference on Circuits, Communication, Control and Computing, Bangalore,2014,pp.452-454.

[28] R. Devika, S. V. Avilala and V. Subramaniyaswamy, "Comparative Study of Classifier for Chronic Kidney Disease prediction using Naive Bayes, KNN and Random Forest," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2019, pp. 679-684.

[29] G. Kaur and A. Sharma, "Predict chronic kidney disease using data mining algorithms in hadoop," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 973-979.

[31] Kamalesh M.D., Predictind the risk of diabetes mellitus to subpopulations using association rule mining, Proceedings of the International Conference of SoftComputing systems, Advances in Intelligent Systems and Computing col.397, Springer (2016)

[32] Revathy,B.Parvathavarthini,Shiny Caroline, Decision Theory, an Unprecedented Validation Scheme for Rough-Fuzzy Clustering, International Journal on Artificial Intelligence Tools, World Scientific Publishing Company ,Vol.25,No.2,2016

[33] Revathy Subramanion, Parvathavarthini Balasubramanian and Shajunisha Noordeen, Enforcement of Rough Fuzzy Clustering Based on Correlation Analysis,International Arab Journal of Information technology,IAJIT,Vol 14,No 1,91-98,2017.

[34] P. Panwong and N. Iam-On, "Predicting transitional interval of kidney disease stages 3 to 5 using data mining method," 2016 Second Asian Conference on Defence Technology (ACDT), Chiang Mai, 2016, pp. 145-150.

[35] S. Vijayarani, S. Dhayanand, "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS", International Journal of Computing and Business Research (IJCBR), vol. 6, no. 2, 2015.