

Random forest and support vector machine-based hybrid liver disease detection

Tsehay Admassu Assegie¹, Rajkumar Subhashni², Napa Komal Kumar³, Jijendra Prasath Manivannan⁴, Pradeep Duraisamy⁵, Minychil Fentahun Engidaye¹

¹Department of Computer Science, College of Natural and Computational Science, Injibara University, Injibara, Ethiopia

²Department of Computer Science and Applications, St. Peter's Institute of Higher Education and Research, Tamil Nadu, India

³Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, Tamil Nadu, India

⁴Data Analyst, HCL Technologies, Chennai, India

⁵Department of Computer Science and Engineering, M.Kumarasamy College of Engineering, Karur, Tamilnadu, India

Article Info

Article history:

Received Mar 11, 2022

Revised Apr 21, 2022

Accepted May 18, 2022

Keywords:

Cirrhosis

Liver disease detection

Machine learning

Random forest

Support vector machine

ABSTRACT

This study develops an automated liver disease detection system using a support vector machine and random forest detection techniques. These techniques are trained on data containing the information collected from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. The proposed system can detect the presence of liver disease in the test set. The random forest model is used for recursive feature elimination at the pre-processing stage and the support vector machine is trained on the optimal feature set. The experimental result shows that the proposed support vector machine (SVM) model has achieved 78.3% accuracy.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Tsehay Admassu

Department of Computer Science, Injibara University

P.O.B: 40, Injibara, Ethiopia

Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

There has been a significant increase in the number of liver disease patients over the last few years, resulting in many cases of death every year. Liver disease is one of the leading causes of mortality worldwide and constitutes a wide range of diseases with varied or unknown etiologies. For instance, a study shows that in 2017 1.32 million deaths worldwide, or 2 to 4% of all annual deaths were caused directly due to cirrhosis [1]-[3]. With the help of automated decision-making methods using a machine learning model, the death caused due to liver disease can be reduced.

Cirrhosis is any factor that harms the liver causing the liver to unfit for its proper functioning. Designing and developing a system for the prediction and diagnosis of liver disease can help doctors or health experts in detecting liver disease correctly [4]. The main purpose of the liver disease prediction system is to classify the given observation as a liver disease patient or not liver disease patient based on the symptoms of liver disease used in training a classification algorithm.

Machine learning (ML) is the branch of artificial intelligence playing a major role in healthcare for the diagnosis of diseases [5]. The implementation of an automated liver disease diagnosis system plays an important role in reducing the mortality rate due to liver disease disorder. ML techniques improve the decision-making process by reducing the false positive rate and increasing the true positive rate during liver disease identification.

The major problem in liver disease diagnosis using ML techniques is improving the accuracy of ML algorithms employed in liver disease diagnosis [6]. Improved accuracy results in better diagnosis results reducing the false negatives and finally increasing the precision in the diagnosis of the liver. In disease diagnosis using ML techniques, clinical liver disease symptoms are usually used to identify patterns in the dataset to the class label. The patient will undergo further medical tests if the ML model identifies any pattern matching the positive class label. However, all the symptoms used for pattern matching that are presented in the liver disease dataset have no importance in the ML learning process. Hence, identifying the features that better characterize liver disease prediction is paramount in developing a more accurate liver diagnosis and prediction model [7]. Hence, this study is devoted to answering the following questions:

- a. What are the risk factors for liver disease?
- b. What are the liver disease features that have higher importance to the learning process of the support vector machine?
- c. How to improve the predictive accuracy of the support vector machine?

The rest of this work is organized as follows: section 2 presents related work on liver disease diagnosis by using various machine learning algorithms. Section 3 discusses the method and dataset used for simulation and experimentation. Section 4 presents the results achieved, and finally, section 5 concludes the work.

2. RELATED WORK

Several studies have been conducted to predict liver disorders using various machine learning algorithms [8]. A framework for liver disease prediction is developed using a clinical liver disease dataset. The developed framework is implemented using hybrid feature selection and regression analysis. The model achieved 89.21% for liver disease diagnosis.

Hashem and Mabrouk [9], developed a decision tree model for the prediction of the normal early stages of cirrhosis stages of the patient. Moreover, the study compares the decision tree model with the random forest (RF) model and the simulation result shows that higher accuracy of 70.67% is achieved with the random forest model as compared to the decision tree model. A support vector machine (SVM) algorithm-based liver disease diagnosis model is proposed in [10]. The SVM is trained on a clinical liver disease dataset collected from the University of Irvine UCI machine learning repository. The result of the experiment shows that promising result is obtained in liver disease prediction using the developed SVM model. The prediction accuracy achieved by the SVM model is 73.2% using the UCI data repository.

In addition, Afrin *et al.* [11] conducted a comparative study on the prediction of liver disease using SVM and an Adaptive boosting algorithm. The SVM and an adoptive boosting algorithm are trained using 583 samples of liver disease dataset collected from the University of California Irvine (UCI) data dataset. The simulation result shows that adaptive boosting outperforms the SVM model with a prediction accuracy of 74.65%. Similarly, in [12], [13], a comparative study is conducted to analyze the performance of K-nearest neighbor (KNN), random forest, decision tree, and adoptive boosting algorithm using the UCI liver disease dataset. The result shows that the decision tree model outperforms as compared to KNN, random forest, and adaptive boosting algorithm for liver disease prediction.

Geethaet and Arunachalam [14], conducted comparative study on four machine learning algorithms namely, random forest, logistic regression, artificial neural network (ANN), and Naïve Bayes (NB) is conducted. In the experimentation for comparative analysis of the performance of the four algorithms, accuracy is used as a criterion for measuring the performance. The result shows that random forest outperforms with the highest accuracy of 84.29% compared to the logistic regression, ANN, and NB. A comparative study on logistic regression and SVM for liver disease prediction shows that SVM outperforms as compared to logistic regression. Liver disease prediction accuracy of 75.04% is achieved using SVM.

In [15] and [16], applied KNN and NB to develop a liver disease prediction model. The prediction accuracy for the KNN model achieved 72.5% while the NB model achieved a prediction accuracy of 63.19%. Hence, the KNN model outperforms as compared to the NB model for liver disease prediction.

Other studies [17]-[19] show that feature selection is important for improving the performance of the machine learning model for liver disease diagnosis. With feature selection, the overlapping symptoms of a disease used in the training of the machine learning process can be interleaved. Hence, based on the literature, the researchers decided to apply recursive feature elimination for determining the significant features for improved performance. Moreover, SVM is employed for model training because different studies [20]-[25] show that the SVM model is effective in multi-class classification tasks such as liver disease prediction.

As shown in the literature survey (section 2), the previous studies do not consider the effect of class imbalance on the predicted accuracy. Although better accuracy is achieved by the prior works the accuracy does not provide the class-wise performance of an ML model on liver disease diagnosis. Moreover, the features that contribute more to the learning process of the different machine learning algorithm is not presented in the literature. Thus, this work aims to address this gap i.e., to investigate liver disease diagnosis feature that is

relevant to the learning process of support vector machine and balancing the dataset using the synthetic minority oversampling technique (SMOTE).

3. METHOD

To select the optimal feature subset for improved performance on the dialogists of liver disease the following steps are followed as demonstrated in Figure 1. The proposed model is developed using the random forest for selecting convenient feature attributes of liver disease by calculating the feature importance and ranking the features according to their importance. Then the selected optimal feature subset is given as input to the SVM and the SVM is trained using the feature subset selected by the random forest feature ranking method.

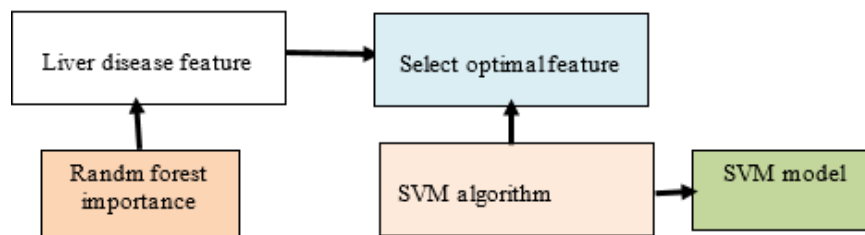


Figure 1. SVM and RF-based hybrid model for liver disease detection

3.1. Liver disease dataset analysis

The liver disease dataset contains the data collected from the Mayo Clinic trial in major PBC of the liver conducted for ten years between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met permissibility conditions for the randomized placebo-controlled test of the drug D-penicillamine. The first 312 cases in the dataset participated in the randomized trial and contain largely complete data. The additional 112 cases did not contribute to the clinical test but subscribed to having basic measurements documented and to be followed for persistence. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants. The proposed SVM-based liver disease diagnosis model is trained on the liver disease features shown in Table 1.

Table 1. The variables in the liver disease dataset

Variables of liver disease dataset	
1.	Patient-ID
2.	Number of days between registration and the earlier of death, transplantation, or study analysis
3.	Status: status of the patient C=censored, CL=censored due to liver, or D=death
4.	Drug: type of drug D-penicillamine or placebo
5.	Age: age of the patient in days
6.	Sex: M=male or F=female
7.	Ascites: presence of ascites N=No or Y=Yes
8.	Hepatomegaly: hepatomegaly, not present=N or present=Y
9.	Spiders: the condition of spiders, no spider=N or spider present=Y
10.	Edema: the condition of edema, no edema, and no diuretic therapy for edema=N, S=edema present with no diuretics, or edema determined by diuretics, or Y=edema withstanding diuretic therapy.
11.	Bilirubin: serum bilirubin in mg/dl
12.	Cholesterol: serum cholesterol in mg/dl
13.	Albumin: albumin in gm/dl
14.	Copper: the value of urine copper in ug/day
15.	Alk Phos: the value of alkaline phosphatase in U/litter
16.	SGOT: the value of SGOT in U/ml
17.	Triglycerides: triglycerides in [mg/dl]
18.	Platelets: platelets per cubic [ml/1000]
19.	Prothrombin: prothrombin time in seconds [s]
20.	Stage: histologic stage of disease (1, 2, 3, or 4)

To explore the effect of a feature on the SVM model, the authors trained the model on the original input feature, and the RFECV feature selection is applied to select the optimal feature. After that, the model is

trained in the optimal feature subset returned by the RFECV feature selection method. To quantify the effect of feature selection, accuracy is used for measuring the performance. The accuracy for SVM is determined by using the formula given in (1),

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} * 100 \tag{1}$$

where TP is true positive, the observations predicted as liver disease patients that belong to a patient class, TN is true negative, the observation predicted as liver disease negative patients that belong to the not disease class. FP is false positive, the observation predicted as liver disease patient that belongs to liver disease negative class, and FN is false negative, the observations predicted as liver disease negative but belong to the liver disease patient class.

4. RESULTS AND DISCUSSION

This section presents the features that are important for training the SVM model selected using random forest-based feature elimination. The effect of the irrelevant feature and the variation in accuracy using the optimal feature and original input feature is analyzed.

4.1. Feature importance analysis

Figure 2 demonstrates the significant features that are relevant to identifying the true positive instances of liver disease. As demonstrated in Figure 2, features such as aspartate aminotransferase, alkaline phosphate, total bilirubin, direct bilirubin, albumin, and age have a significant effect on the model output, in contrast, the albumin and globulin ratio, total proteins and gender has a lower impact on the model output. Thus, the use of the features with a higher impact on model output improves the performance of the SVM model for liver disease prediction.

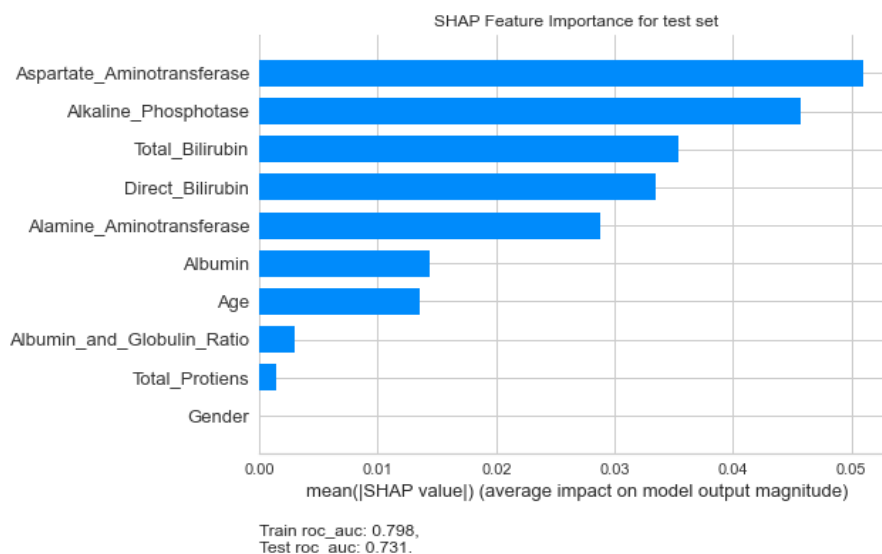


Figure 2. SHAP summary plot of liver disease features

Figure 3 demonstrates the permutation feature importance using random forest-based feature elimination. As demonstrated in Figure 3, the features selected by recursive feature elimination using the random forest model are aspartate aminotransferase, alkaline phosphate, amine, aminotransferase, age, and total bilirubin, albumin, and direct bilirubin. Thus, the SVM model is trained on these features and the performance when the SVM is trained on the features shown in Figure 3 is demonstrated in Table 1. The recursive feature elimination with cross-validation (RFECV) random forest is employed to determine the optimal number of input features. As demonstrated in Figure 3, the RFECV returns 8 features as optimal input features that produce the highest possible accuracy of 77.5%. Based on the number of input features determined by the RFECV, the first 8 important features are selected by the permutation-based feature importance shown in Figure 2. Then the model is trained on the 8 optimal input features to obtain higher accuracy.

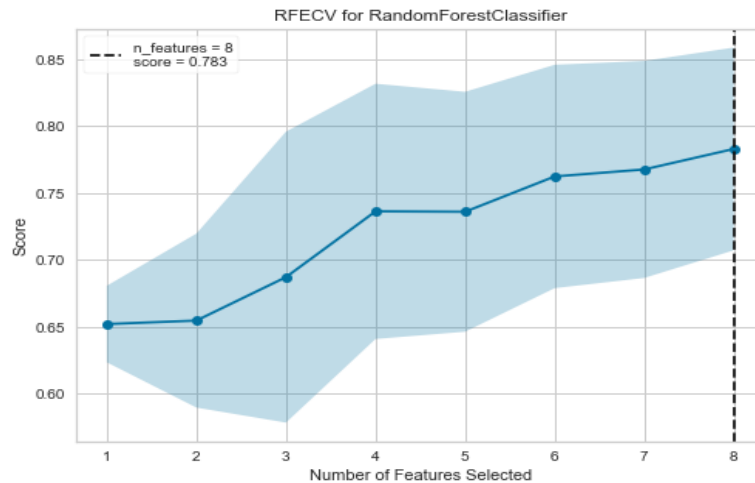


Figure 3. Permutation feature importance using random forest

5. CONCLUSION

This work proposed an automated liver disease detection system by using a random forest model. The developed system is tested on a real-world liver disease dataset. Simulation using the test set shows that the developed system has achieved acceptable accuracy. With such an automated system the required precision can be achieved as the system aids in the decision-making process of liver disease diagnosis. The simulation shows that the developed liver disease diagnosis system has 78.3% accuracy. The result also shows that the accuracy improves by 10.2% when the RFECV is used for feature selection using synthetic minority oversampling. Thus, the developed model is significantly important to assist the medical expert in the prediction of liver disease.

ACKNOWLEDGEMENTS

The authors would like to thank Injibara University for providing a laptop for conducting the simulations in this work.





REFERENCES

- [1] J. Singh, S. Bagga, and R. Kaur, "Software-based Prediction of Liver Disease with Feature Selection and Classification Techniques," *Procedia Computer Science*, vol. 167, pp. 1970–1980, 2020, doi: 10.1016/j.procs.2020.03.226.
- [2] W.-J. Kim *et al.*, "Development and validation of a novel scoring system for the prediction of disease recurrence following resection of colorectal liver metastasis," *Asian J. of Surg.*, vol. 43, no. 2, pp. 438–446, 2020, doi: <https://doi.org/10.1016/j.asjsur.2019.06.001>.
- [3] W.-K. Seto and M. S. Mandell, "Chronic liver disease: Global perspectives and future challenges to delivering quality health care," *PLOS ONE*, vol. 16, no. 1, 2021, doi: <https://doi.org/10.1371/journal.pone.0243607>.
- [4] A. D. Vincentis *et al.*, "A Polygenic Risk Score to Refine Risk Stratification and Prediction for Severe Liver Disease by Clinical Fibrosis Scores," *Clin. Gastroenter. and Hemat.*, vol. 20, no. 3, pp. 658–673, 2021, doi: <https://doi.org/10.1016/j.cgh.2021.05.056>.
- [5] N. Jiang, Z. Zhao and P. Xu, "Predictive Analysis and Evaluation Model of Chronic Liver Disease Based on BP Neural Network with Improved Ant Colony Algorithm," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–7, 2021, doi:10.1155/2021/3927551.
- [6] B. Sumathy *et al.*, "A Liver Damage Prediction Using Partial Differential Segmentation with Improved Convolutional Neural Network," *Journal of Healthcare Engineering*, vol. 2022, pp. 1–14, 2022, doi:10.1155/2022/4055491.
- [7] M. Mehmood, N. Alshammari, S. A. Alanazi, and F. Ahmad, "Systematic Framework to Predict Early-Stage Liver Carcinoma Using Hybrid of Feature Selection Techniques and Regression Techniques," *Complexity*, vol. 2022, pp. 1–11, 2022, doi:10.1155/2022/7816200.
- [8] N. Nahar and F. Ara, "Liver Disease Prediction by Using Different Decision Tree Techniques," *International Journal of Data Mining & Knowledge Management Process*, vol. 8, no. 2, pp. 1–9, 2018, doi: 10.5121/ijdkp.2018.8201.
- [9] E. M. Hashem and M. S. Mabrouk, "A Study of Support Vector Machine Algorithm for Liver Disease Diagnosis," *American Journal of Intelligent Systems*, vol. 4, no.1, pp. 9–14, 2014, doi: 10.5923/j.ajis.20140401.02.
- [10] D. Devikanniga, A. Ramu, and A. Haldorai, "Efficient Diagnosis of Liver Disease using Support Vector Machine Optimized with Crows Search Algorithm," *EAI Endorsed Transactions on Energy Web*, vol. 20, no. 29, 2020, doi: 10.4108/eai.13-7-2018.164177.
- [11] S. Afrin *et al.*, "Supervised machine learning based liver disease prediction approach with LASSO feature selection," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, pp. 3369–3376, 2021, doi: 10.11591/eei.v10i6.3242.
- [12] R. A. Khan, Y. Luo, and F.-X. Wu, "Machine learning based liver disease diagnosis: A systematic review," *Neurocomputing*, vol. 468, pp. 492–509, 2021, doi: 10.1016/j.neucom.2021.08.138.
- [13] C.-C. Wu *et al.*, "Prediction of fatty liver disease using machine learning algorithms," *Computer Methods and Programs in Biomedicine*, vol. 170, pp. 23–29, 2019, doi:10.1016/j.cmpb.2018.12.032.
- [14] C. Geethaet and A. R. Arunachalam, "Evaluation based Approaches for Liver Disease Prediction using Machine Learning





- Algorithms,” *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, doi: 10.1109/ICCCI50826.2021.9402463.
- [15] H. Hartatik, M. B. Tamam, and A. Setyanto, “Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms,” *2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS)*, 2020, doi: 10.1109/ICORIS50180.2020.9320797.
- [16] F. Mostafa, E. Hasan, M. Williamson, and H. Khan, “Statistical Machine Learning Approaches to Liver Disease Prediction,” *Livers*, vol. 1, no. 4, 294–312, 2021, doi: 10.3390/livers1040023.
- [17] S. J. Sushma, T. A. Assegie, D. C. Vinutha, and S. Padmashree, “An improved feature selection approach for chronic heart disease detection,” *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 6, December 2021, pp. 3501–3506, doi: 10.11591/eei.v10i6.3001.
- [18] M. H. Arif, A.-R. Hedar, T. H. A. Hamid, and Y. B. Mahdy, “SS-SVM (3SVM): A New Classification Method for Hepatitis Disease Diagnosis,” *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 4, no. 2, pp. 53–58, 2013, doi: 10.14569/IJACSA.2013.040208.
- [19] N. Razali, A. Mustapha, M. H. Abd Wahab, S. A. Mostafa, and S. K. Rostam, “A Data Mining Approach to Prediction of Liver Diseases,” *Journal of Physics: onference Series*, vol. 1529, pp. 1–7, 2020, doi:10.1088/1742-6596/1529/3/032002.
- [20] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, “A Systematic Machine Learning-Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression,” *Scientific Reports*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-018-20166-x
- [21] C.-L. Liu, R.-S. Soong, W.-C. Lee, G.-W. Jiang, and Y.-C. Lin, “Predicting Short-term Survival after Liver Transplantation using Machine Learning,” *Scientific Reports*, vol. 10, pp. 1–10, 2020, doi: 10.1038/s41598-020-62387-z.
- [22] S. Ambesange et al., “Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques,” *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, 2020, doi: 10.1109/CCEM50674.2020.00030.
- [23] K. Thirunavukkarasu, A. S. Singh, Md Irfan, and A. Chowdhury, “Prediction of Liver Disease using Classification Algorithms,” *2018 4th Int. Conference on Computing Communication and Automation (ICCCA)*, 2019, doi:10.1109/CCAA.2018.8777655.
- [24] T. A. Assegie, “Support Vector Machine and K-nearest Neighbor Based Liver Disease Classification Model,” *Indonesian J. of Electronics Electromedical Engineering and Medical Informatics*, vol. 3, no. 1, pp. 9–14, Nov. 2020, doi: 10.1234/jeeemi.v1i1.9xx.
- [25] J. H. Joloudari, H. Saadatfar, A. Dehzangi, and S. Shamshirband, “Computer-aided decision-making for predicting liver disease using PSO-based optimized SVM with feature selection,” *Informatics in Medicine Unlocked*, vol. 17, 2019, doi:10.1016/j.imu.2019.100255.

BIOGRAPHIES OF AUTHORS







Tsehay Admassu Assegie     obtained a Master of Science degree in Computer Science from Andhra University Faculty of Science, India 2016. He received B.Sc. (Computer Science) from Dilla University, Ethiopia in 2013. His research interest includes machine learning, data mining, bioinformatics, network security, and software-defined networking. He has published over 33 journal articles in international journals and conferences. He can be contacted at email: tsehayadmassu2006@gmail.com.







Rajkumar Subhashni     is currently working as Assoc. Professor in the Department of Computer Science and Applications, St. Peter’s Institute of Higher Education and Research. Her research area includes software engineering, machine learning, and artificial intelligence. She can be contacted at email: subhashniraj2018@gmail.com.







Napa Komal Kumar     is currently working as an Assistant Professor in the Department of Computer Science and Engineering at St. Peter’s Institute of Higher Education and Research, Avadi, Chennai. His research interests include machine learning, data mining, and cloud computing. He can be contacted at email: komalkumarna@gmail.com.







Jijendra Prasath Manivannan     is currently working as Data Analyst at HCL TECHNOLOGIES, Chennai, TamilNadu, India. His research interest includes satellite imagery, remote sensing, nasal sensing, data engineering, machine learning, image processing, and natural language processing. He can be contacted at email: jijendiran1999@gmail.com.



Pradeep Duraisamy     is currently working as an Assistant Professor of Computer Science and Engineering, M.Kumarasamy College of Engineering, Karur, Tamilnadu, India. He has received his Ph.D. from Anna University, Chennai, Tamilnadu, India. He has More than 10 years of academic experience in engineering colleges. He has published around 25 research articles in various Journals, Conferences, Book Chapters, and Patents. He has organized around 22 events for student and faculty communities. He is a member of ISTE and IAENG. His area of interest includes cyber security, artificial intelligence, data mining, big data, and IoT. He is contributing as a reviewer of various reputed journals. He can be contacted at email: pradeepdurai.vdr@gmail.com.



Minychil Fentahun Engidaye     is currently working as a lecturer in the Department of Computer Science, Injibara University, Injibara, Ethiopia. His research interest includes machine learning and natural language processing. He can be contacted at email: minychil@gmail.com.