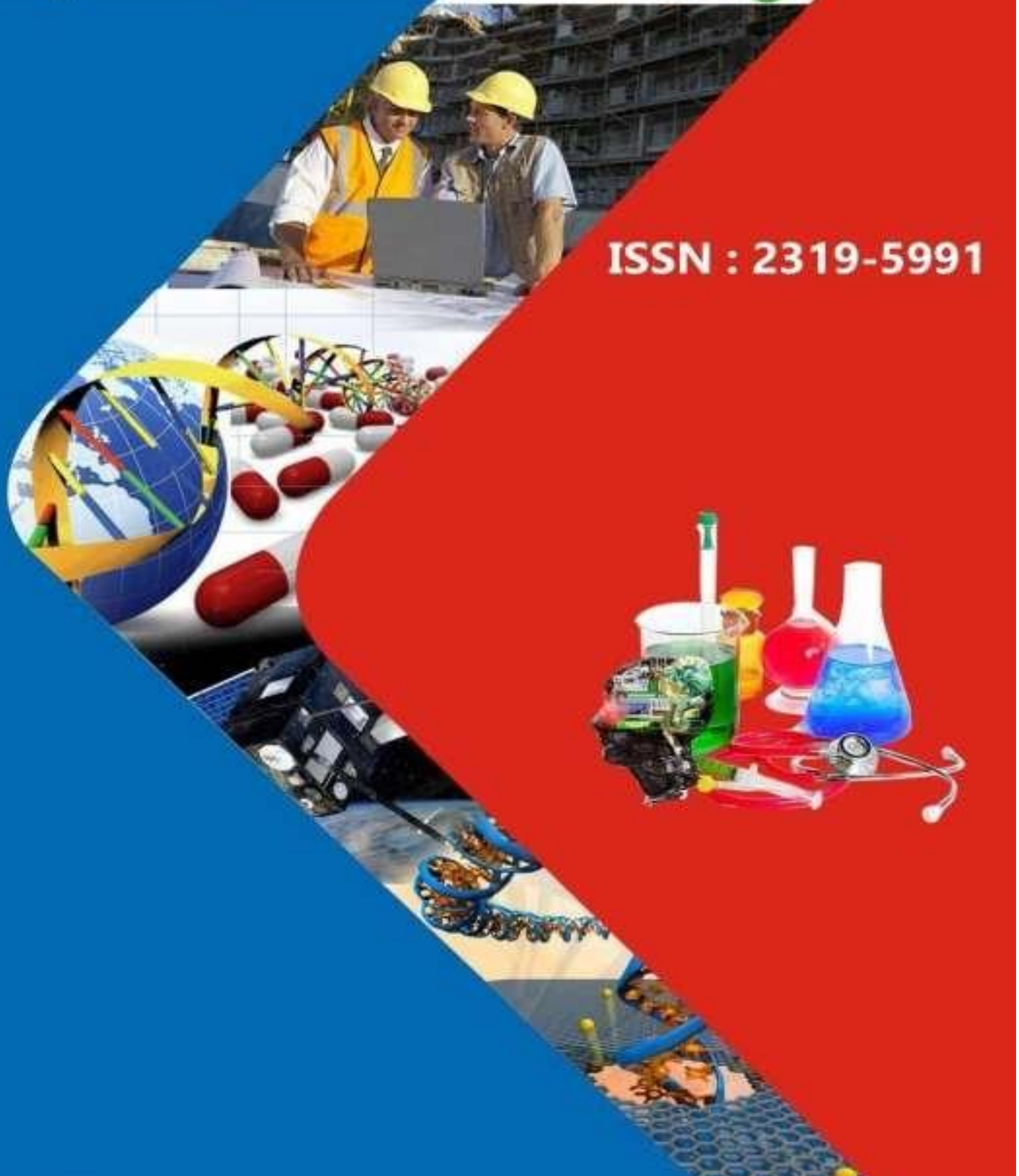
IJERST

International Journal of Engineering Research and Science & Technology



ISSN : 2319-5991

www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue.2, April 2024

RANDOM FOREST OPTIMISING LIVER DISEASE PREDICTION THROUGH DIFFERENT DATA BALANCING TECHNIQUES

Kapilavai Purushotham Raju
Master of Engineering Student
Department Of ECE
University College of engineering Osmania University

ABSTRACT

A prevalent condition, apart from heart attacks, which take many lives, is liver disease. Due to the fact that liver disease is usually detected late and causes mortality. The number of liver patients is increasing due to a variety of circumstances, such as drinking contaminated water, inhaling toxic gas, abusing alcohol excessively, and other behaviours that might affect health parameters. By using machine learning prediction models, these health parameters may be utilised to predict liver ailments in their early stages. The machine-learning model in this study is developed using the Indian Liver Patient Dataset (ILPD), which is based on Indian patients. Several preprocessing methods are used with the Random Forest (RF) algorithm to forecast the illness. The data collection is examined using univariate and bivariate analysis to check for imbalance, skewness, and outliers. After that, appropriate algorithms are used to remove outliers, and different techniques for oversampling and undersampling are applied to balance the data. Grid search and feature selection are used in conjunction with hyper parameter adjustment to further refine the model. The final model has 100% accuracy and performs very well across all criteria.

I. INTRODUCTION

An essential part of the human body, the liver is found in the right upper abdomen just below the rib cage. It keeps the body's blood sugar levels in a healthy range and eliminates toxins from the body. Although the body's organs are capable of self-healing, excessive alcohol use and exposure to polluted air and water harm the liver, which increases the risk of liver failure. The answer is liver transplantation, although it is more expensive and has a lower success rate. Early detection of liver disease can lower the risk of liver failure. Based on a data set that combines important health indicators for both people with and without disorders, the machine-learning algorithm can predict diseases. An efficient data set with accurate illness classification representation is required for model construction. The Indian Liver Patient Dataset, available at ics.uci.edu, is used in this study. The classification of liver illnesses is possible thanks to a variety of machine learning methods. The study takes a step-by-step approach to how to modify the ML module for the Random Forest method rather than choosing the algorithm that performs best. The Random Forest approach is used to develop a model since the model is not optimised for highly particular data because it is trained on multiple samples of data collected by splitting data. The paper's major objective is to thoroughly examine how an unbalanced data set allows models to be modified beyond a point of saturation. Later parts include a summary of the various balancing approaches explored and their effects on performance. The literature review is included in Section 2 of this paper. Several models were developed iteratively, with each critical step involving the data set and its effects on performance optimisation being highlighted. This is covered in depth, including with figures and results, in the outcome and model development section. The final portion contains the conclusions.

Purpose:

ISSN 2319-5991 www.ijerst.com

Vol. 17, Issue.2, April 2024

1. **Enhancing Healthcare Accuracy**: Increasing the precision of models for predicting liver disease to aid in early identification and treatment.
2. **Reducing Errors**: Reducing false positives and false negatives in the diagnosis of liver illness.
3. **Generalizing Model Performance**: Ensuring that the model functions effectively across a range of patient demographics.
4. **Addressing Healthcare Disparities**: Fostering fair medical results for all racial and ethnic groupings.
5. **Cost-Efficiency**: By avoiding pointless testing and interventions, healthcare resources can be conserved..
6. **Advancing Research**: Making a contribution to the field of machine learning and healthcare analytics.
7. **Patient Engagement**: Encouraging people to participate actively in the management of their healthcare.

Motivation:

1. **Healthcare Impact**: Potential for liver disease sufferers to live longer and have better quality of lives.
2. **Data Imbalance Challenge**: The problem of unbalanced datasets in medical diagnosis is addressed.
3. **Machine Learning Enhancement**: Enhancing the predictive power of the Random Forest algorithm by using data balancing methods.
4. **Equity and Inclusivity**: Ensuring that all patients have access to efficient healthcare options.
5. **Economic Benefits**: Lowering medical expenditures by making more accurate predictions.
6. **Research Contribution**: Advancing knowledge at the nexus of machine learning and healthcare.
7. **Patient Empowerment**: Enabling people to take active measures to manage their health.

II. LITERATURE SURVEY

1. A. Sharma and D. Rani's "Machine Learning Techniques for Liver Disease Prediction: A Review" (2020)

• This review article offers insights into different machine learning methods for predicting liver illness, including solutions for data imbalance.

2. "Addressing Class Imbalance in Medical Datasets" (2019) by J. Fernández-Navarro et al.

• This study examines methods for dealing with class imbalance in medical datasets, which is an important factor in raising the accuracy of liver disease prediction.

3. L. Breiman's "Random Forest: A Review" from 2016.

• The foundational work of L. Breiman examines the Random Forest algorithm, a key element of the project, and its applicability in a variety of fields, including healthcare.

4. I. Rajkomar, et al., "Machine Learning for Healthcare:

On the Verge of a Major Shift in Healthcare Epidemiology" (2018) This article examines the possible effects of machine learning on healthcare, highlighting the significance of precise prediction models in enhancing patient outcomes.

5. "Machine Learning and Data Mining Methods in Diabetes Research" (2020) by L. Wei and et al.

• This study explores the application of machine learning and data balancing techniques in medical research and their potential to improve disease prediction, with a particular focus on diabetes.

6. N. V. Chawla, et al., "Overcoming Class Imbalance in Medical Datasets: A Review of Techniques" (2017)

This thorough review study especially addresses the topic of class imbalance in medical datasets and explores numerous ways for handling it successfully.

7. "Enhancing Predictive Accuracy in Imbalanced Datasets" (2018) by P. B. Abdi

• This study investigates techniques for enhancing predictive accuracy in imbalanced datasets, offering insights that can be used to improve models for predicting liver disease.

8. "Ethical Considerations in Machine Learning for Healthcare" (2021) by E. K. Garcia et al.

• This article explores ethical issues surrounding machine learning in healthcare with a focus on justice and fairness in model predictions, which is in line with the project's objectives.

9. "Patient-Centered Machine Learning in Healthcare" (2019) by M. E. Rogers et al.

This article examines the value of patient-centered techniques in healthcare machine learning projects, which is pertinent to the project's objective of patient engagement in the prediction of liver disease.

2.1 FEASIBILITY STUDY :- The feasibility of the project and the likelihood that consumers will find the application useful are both looked at in the preliminary research. The main objective of the feasibility study is to test the technical, operational, and financial feasibility of adding new modules, debugging traditional desktop-centric applications, and porting them to mobile devices. All systems are conceivable if given an infinite amount of time and resources. The following components make up the feasibility study for the preliminary investigation: These three crucial aspects should receive specific consideration when determining the project's viability:

♣ Technical Feasibility

♣ Economic Feasibility

♣ Operational Feasibility

2.1.1 TECHNICAL FEASIBILITY

Evaluating the technical feasibility is the most difficult part of a feasibility study. This is because there are now not enough comprehensive system designs, which makes it difficult to access problems with performance, costs (due to the kind of technology to be employed), etc. A number of elements need to be considered while conducting a technical analysis. 1. Be knowledgeable about the various technologies incorporated into the proposed system. Before we begin the project, we must have a very clear understanding of the technologies that will be required for the development of the new system. 2. Ascertain whether the business now possesses the required technologies: Page | 6 Does the organisation have the required technology? Does the capacity match the needs in that case? Question: "Will the current printer be able to handle the new reports and forms required for the new system?"

2.1.2 ECONOMICFEASIBILITY

Economic feasibility tries to strike a balance between the advantages of having the new system in place and the costs of designing and putting the new system into place. Thanks to this feasibility assessment, the senior management has an economic justification for the new system. In this case, a plain economic study with a precise cost-benefit analysis would be far more helpful. stronger customer satisfaction, better product quality, faster information retrieval, better decision-making, accelerated activities, improved process accuracy, better documentation and record keeping, and stronger workforce morale are a few examples of these.

2.1.3 OPERATIONALFEASIBILITY

Proposed concepts are only useful if they can be developed into information systems that meet the operational requirements of the company. Simply expressed, the purpose of this feasibility test is to determine whether the system will work once it has been developed and installed. Exist any significant implementation barriers? To help decide

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue.2, April 2024

whether a project is operationally feasible, ask the following questions: Is the project receiving enough management and user support? If the current system is so well-liked and widely used that people have a hard time understanding why it needs to change, there may be pushback. The user accepts the present business practises. Users may accept a change if it leads to a more helpful and functional system even if they don't now. Has the user participated in the project's creation and planning? Early involvement increases the likelihood that the project will be successful and reduces the likelihood of system opposition in general. Considering that the proposed approach was designed to make things easier.Under the current manual system, the new strategy was believed to be operationally workable.

III. SYSTEM ANALYSIS AND DESIGN

Existing System:

Description of the Existing System:

The current system for predicting liver disease typically uses machine learning algorithms or conventional statistical models. These systems predict the presence or absence of liver illnesses based on past patient data, clinical indicators, and laboratory test findings. However, when dealing with imbalanced datasets, they frequently lack resilience.

Disadvantages of the Existing System:

1. **Limited Accuracy**: The current system may be inaccurate, especially when dealing with unbalanced datasets, which could result in a high percentage of false positives or false negatives for diagnosing liver illness.
2. **Data Imbalance**: These systems frequently have trouble solving problems with data imbalance. They might have a bias in favour of the majority class, which would make their forecasts about the minority class erroneous.
3. **Outdated Algorithms**: The predictive capability of many existing systems is constrained by the usage of antiquated algorithms that do not fully utilise the promise of contemporary machine learning techniques.

Proposed System:

Description of the Proposed System:

The suggested system uses the Random Forest algorithm and several data balancing strategies to improve the prediction of liver illness. To dramatically increase prediction accuracy, this sophisticated system combines the benefits of Random Forest with data preparation techniques.

Advantages of the Proposed System:

1. **Improved Accuracy**: The suggested approach is anticipated to have noticeably higher accuracy in predicting liver illness. It solves the shortcomings of the current system and provides more trustworthy diagnoses by utilising Random Forest and data balancing approaches.
2. **Balanced Data**: Techniques for data balance, like oversampling and undersampling, are incorporated into the suggested system. This lessens the effect of data imbalance and guarantees accurate and impartial predictions.
3. **Flexibility and Scalability**: The suggested solution is made to be adaptable and expandable. It can include updated algorithms, respond to changes in healthcare data, and efficiently handle rising patient record volumes.

3.2 REQUIREMENTS SPECIFICATION

HARDWARE REQUIREMENTS

- ➢ System : Any Latest Processor
- ➢ Hard Disk : 500 GB.
- ➢ Monitor : Any LED / LCD
- ➢ Mouse : Optical Mouse.
- ➢ RAM : 8 GB (Min)

SOFTWARE REQUIREMENTS

- Operating system : Windows 7 / 8 / 10 (64-bit)
- Coding Language : Python 3.7.0
- Front-End : Python
- UI : Tkinter
- Data Base : MySQL.

3.4 SYSTEM DESIGN

System design is the process or art of defining a system's architecture, components, modules, interfaces, and data in order to satisfy preset requirements. It might be viewed as a systems theory application to the process of product development. There are certain overlaps and synergies between the domains of systems analysis, systems architecture, and systems engineering.

3.4.1 SYSTEM ARCHITECTURE

To discover brain tumours, a thorough computer-aided diagnosis approach is developed. Clinicians can swiftly and accurately identify the cancerous cells by using computer-aided diagnostics (CAD). This method mainly concentrates on figuring out the stage of brain cancer. The proposed approach for the detection of brain tumours involves three key steps: preprocessing, feature extraction, and classification.

Preprocessing is done on the acquired MRI scan image, as was already mentioned. The segmented image is then used to extract the features. Finally, we classify the image based on the retrieved region and features. With our suggested solution, we have made an effort to address the problems we have had with the current system.

This technique can reveal if a tumour is malignant or not. If a tumour is malignant, cancer will grow as a result. If the tumour is not cancerous, the results are No Tumour.

This evidence suggests that the cancer is curable with the appropriate medical care. Consequently, the patient's tumour might be manageable even at an early age.

3.4.2 Input Design

The input design phase of the software development life cycle is extremely important, and developers must pay great attention to it. The input design's objective is to give the application the most accurate data possible. Inputs must therefore be properly prepared in order to minimise feeding errors. Software engineering principles dictate that the input forms or screens be made with a validation control on the input limit, range, and other related factors.

Validations.

Nearly all of this system's modules have input screens. Error messages are intended to alert the user when mistakes are made and guide him appropriately to stop the creation of valid entries. Let's take a closer look at this under module design. Input design is the process of converting user-generated data into a computer-based representation. The input design seeks to guarantee reasonable and error-free data entry. The error in the input is controlled by the input design. The application's simplicity of use was a design priority. The forms have been designed such that the cursor will be in the proper place for input when they are processed. The user may occasionally be offered the option to select an appropriate input from a range of options related to the field. Validation is required for each entered piece of data.

When a user enters incorrect information, an error message is displayed, and once they have completed all the entries on the current page, they are given the option to move on to the subsequent pages.

3.4.3 Output Design

In order to create a productive internal communication system for the company, notably between the project manager and his team, or, to put it another way, between the administrator and the customer, the output from the computer is mostly required. The ultimate result of VPN is a system that allows the project manager to manage his clients by adding new clients and giving them new projects, monitoring the feasibility of the projects, and providing each client with, based on the project given to him, user-level access to particular folders. Once a project is completed, the client may be given a new one. The initial stages itself uphold user authentication systems. However, only the administrator has the power to assign projects to new users and confirm their registration. A new user can be created by either the administrator or a user.

The application starts to run when it is initially executed. The server needs to be launched. A local area network will be used for the project's execution, enabling the server machine to act as the administrator while the other linked systems can act as clients. The developed system is quite user-friendly and easy enough for someone using it for the first time to comprehend.

IV. CONCLUSION

An RF algorithm is used to make predictions after a range of preprocessing methods are applied to the imbalanced data in this study. If the data set is uneven, then just pre-processing measures like replacing missing values, handling outliers, and converting the data set will not enhance the results. Performance is improved utilising all optimisation strategies, such as feature selection, hyperparameter modification, and PCA, up to the value shown in the result table. To further refine a model, the data set has to be balanced, which is accomplished in this study by using a number of oversampling and undersampling techniques. It has been discovered that the prediction accuracy of an oversampled dataset is poorer than that of an undersampled dataset; this suggests that oversampling increases result variance and degrades the correlation and relationship between characteristics and the target label. It also demonstrates that clean data is essential for building efficient models and that more data does not necessarily equate to better results. We may use the same methods on a new set of data in the future to evaluate how accurate the predictions are.

REFERENCES

1. https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+D ataset)
2. C. Chao, A. Liaw, and L. Breiman. “ Using random forest to learn imbalanced data.” University of California, Berkeley 110 (2004):
3. 1-12 H. He, Y. Bai, E. A. Garcia, S. Li, “ADASYN: Adaptive synthetic sampling Proceedings of the 5th IEEE International Joint Conference on Neural Networks, pp. 1322-1328, 2008.
4. I. Tomek, “An experiment with the edited nearest-neighbor rule,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 6(6), pp. 448-452, 1976.
5. G. E. A. P. A. Batista, R. C. Prati, M. C. Monard, “A study of the behavior of several methods for balancing machine learning training data,” ACM Sigkdd

ISSN 2319-5991 www.ijerst.com
Vol. 17, Issue.2, April 2024

Explorations Newsletter, vol. 6(1), pp. 20-29, 2004

6. D. Wilson, “Asymptotic Properties of Nearest Neighbor Rules Using Edited Data,” IEEE Transactions on Systems, Man, and Cybernetics, vol. 2(3), pp. 408-421, 1972.
7. J. Laurikkala, “Improving identification of difficult small classes by balancing class distribution,” Proceedings of the 8th Conference on Artificial Intelligence in Medicine in Europe, pp. 63-66, 2001
8. P. E. Hart, “The condensed nearest neighbor rule,” IEEE Transactions on Information Theory, vol. 14(3), pp. 515-516, 1968.
9. Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J & Napolitano, A. “RUSBoost: A hybrid approach to alleviating class imbalance.” IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 40.1 (2010): 185-1