Grenze International Journal of Engineering and Technology, June Issue

Analysis of Recursive Feature and Stochastic Gradient Boosting to Predict Skin Cancer

V. Kakulapati[1], Ch. Shiva Kumar[2], Sd. Faizan[3] and Pritham Kumar[4]
[1-4]Sreenidhi Institute of Science and Technology, Yamnampet, Ghatkesar, Hyderabad, Telangana-501301
Email: vldms@yahoo.com, shivakumar.chinthakunta08@gmail.com, sdfaizan88@gmail.com, prithamkumar10@gmail.com

***Abstract*— The malignancy of the skin is the most prevalent type of cancer in the worldwide population. It makes it easier to find and treat the illness early, which is a crucial aspect in raising the chances of survival for patients. HOwever, Skin Cancer Prediction from Skin Lesions is not a typical problem due to niρh dimension and the number of factors of many redundant or irrelevant features. Conventional procedures for diagnosis and simple machine learning approaches usually take up more computing power, which makes them incorrect. To overcome these problems, this study introduces a machine learning-based skin cancer prediction (SCP) model based on Stochastic Gradient Boosting (SGB) via Recursive Feature Elimination (RFE). Remove the unimportant and unnecessary data to work on important features only. It is called RFE, which is more than minus importance. At this point, we trained these features on a Stochastic Gradient Boosting Model. This trains multiple weak learners, which are then combined in a single strong predictive model that aims to improve the accuracy of predictions. Overall, the introduced system has better prediction in a less complicated manner and provides more reliable results. This will aid healthcare professionals in better detecting skin cancer in the early stages, leading to better decision-making. Clinical data-based cancer prediction is important for facilitating correct diagnosis and treatment.**

***Keywords*— Cancer Prediction, Stochastic Gradient Boosting, Recursive Feature Elimination (RFE), Feature Selection. Machine Learning, Early Detection, Classification, Medical Data Analysis.**

I. INTRODUCTION

A skin cancer is one of multiple forms of cancer around the world and mainly arises with pre-existing proliferating abnormal skin cells after excessive exposure to ultraviolet (UV) rays. Allied detection and analysis provides a window for timely treatment, which is critical to achieve favorable results. However, traditional diagnostic methods depend on clinicians and physical examinations [1], which may lead to delays or human errors. Starting from this leads to a high demand for smart automation systems that will help health professionals detect skin cancer at its initial stage.

With the tremendous rise of Machine Learning (ML) and Artificial Intelligence (AI), predictive analytics has established itself as an extremely useful means to make medical diagnoses. This work proposed Stochastic Gradient Boosting (SGB) [2] with Recursive Feature Elimination (RFE) based Machine Learning approach to predict skin cancer. RFE works by reducing redundant and unnecessary data, allowing you to pinpoint the necessary features with improved efficiency and precision. The last model we used is the Stochastic Gradient Boosting [3] using features that were chosen.

Grenze ID: 01.GIJET.12.2.557_26

The model learns a bunch of weak learners to end up with a superb classifier that gets better prediction performance. Ensure that the pro-posed system makes very fast, low-complexity predictions and produces very reliable output. It helps medical professionals identify it early and encourages better skin cancer diagnosis.

A. Objectives

This research has been primarily interested in the design of a fast, accurate, and non-destructive prediction of skin cancer system using different modes of supervised machine learning. Skin Cancer is probably one of the most well-known diseases for decreasing mortality and improving survival via early detection, especially in the early stages. This prediction system is designed to support medical professionals by providing appropriate and timely predictions.

In this study, we concentrate on the classification algorithm Stochastic Gradient Boosting(SGB). Boosting is an ensemble learning method that combines the power of many weak Learners and trains them to obtain a classifier. This approach enables the classification accuracy to be maximized while minimizing inter-predictor error correlation, as with current machine learning algorithms.

Another importance is to address the issue of high-dimensional medical data. So, one of the problems that is commonly found in most of the medical datasets is irrelevant, redundant, and noisy features, which not only reduces model performance but also increases computational costs. To solve this problem, Recursive Feature Elimination is performed to retain only the important features. This boosts the model's efficiency and gives one more layer of overfitting reduction.

Simultaneously, to enhance computational efficiency by lowering input dimensionality and improving the learning. Leading to fast training and prediction with high accuracy. In addition, in their testing [from a practical perspective], it would also mean that the robustness of the system across datasets would make this appropriate for real-life medical uses as well. Therefore, the system essentially becomes a unifying form of decision-support tool for physicians and medical personnel. Moreover, it can give accurate and repeatable predictions that help in early diagnosis checking, along with reducing human intervention er-rors, which would design better treatment plans leading to increased patient outcome & better healthcare services.

B. Challenges

In most cases, the unbalancing problem appeared when applying to medical datasets (for example, skin cancer prediction). Such an imbalance could lead to bias in the model, lowering posi-tive case detection, which decreases prediction performance. The next big challenge is high-dimensional data. Medical datasets typically contain many features, of which some are likely to be unrelated and/or duplicate information. But it's hard to focus on the most relevant features, as it loses diagnostic infor-mation. Features selection not properly can harm Model accuracy and trust.

Moreover, data preprocessing (it's an extremely important case) presents its own challenges. By defini-tion, the input in machine learning models is medical data that will be used, and there will always be a level of noise, missing values, or inconsistency found in medical data. You only need to hold prepro-cessing (which can be used for cleaning and normalising your data) techniques (not an abuse filter that loses extremely critical information). So, correct pre-processing steps, such as treating data imbalance and selecting only those features suitable for building any skin carcinoma prediction system, are of paramount importance. To overcome the afore-mentioned challenges related to high-dimensional data, a number of feature selection techniques have also been recommended. Therefore, you could keep the most relevant features and remove others with methods like Recursive Feature Elimination (RFE). This allows the model's efficiency to be optimised, while reducing computation cost and increasing prediction accuracy.

Therefore, you could keep the most relevant features and remove others with methods like Recursive Fea-ture Elimination (RFE). This allows the model's efficiency to be optimised, while reducing computation cost and increasing prediction accuracy. But even with those advancements, there are still plenty of sys-tem which are faced with problems of data imbalance, overfitting & low generalisation. Previous models either extract features or classify with both, which is stylistic optimization. Now, it is time to do feature selection using efficient methods together with powerful ensemble algorithms like Stochastic Gradient Boosting: can provide a skin cancer.

II. Background and Purpose

Skin cancer is diagnosed manually, using clinical examination and medical data. It can take a while, and in most cases, human error is involved. Typically, machine learning methods rely on elementary statistical features and simple classifiers [4], which cannot capture the complex patterns in medical datasets. These limitations resulted in the creation of more advanced ensemble methodologies, namely Stochastic Gradi-ent Boosting. It has a better ability to model non-linear relationships in the data, and it also improves clas-sification.

Furthermore, Recursive Feature Elimination (RFE) aids in model optimization by selecting relevant fea-tures and removing noise or redundant features. This study focuses primarily on defining a more reliable predictive system via Stochastic Gradient Boosting boosted with Recursive Feature Elimination to make prediction outcomes better. Due to this combination, skin cancer could be better detected at earlier stages, and more trustworthy decision-making procedures could take place in medical applications.

III. Related Works

Over the past two decades, skin cancer detection has drawn a considerable amount of attention in many countries due to the increase in incidents globally. Detection in the early stages was primarily based on manual examination and conventional image analysis techniques, with dermatologists reviewing skin le-sions through visual inspection. While partial success was obtained with these treatments, they were heav-ily reliant on expert opinion and led to a delayed diagnosis as well as human error.

As machine learning has progressed, different automated approaches for predicting skin cancer have been proposed. Previously, classification tasks have been done by traditional models like Support Vector Ma-chines (SVM), Decision Trees, and Random Forests. While they performed better than hand diagnose but still require the need for extensive feature engineering to work with and outperformed on high-dimensional datasets or replicate features [5]. Image processing techniques are available to extract fea-tures based on skin lesion characteristics, including texture, color, and shape. Although these handcrafted features did provide some useful insights, they often failed to capture complex patterns that exist in medi-cal data. Consequently, when tested on heterogeneous and real-world datasets, their performances were also very limited.

Some deep learning approaches have been popular for this situation in recent years. Image classification has been achieved successfully by Convolutional Neural Networks (CNNs) as CNN-based approaches learn important features automatically [4]. Promising results in binarizing malignant and benign lesions have been achieved using models such as ResNet, Inception Tensor, and DenseNet Cortina [6]. On the other hand, deep learning methods typically need access to not only large amounts of labelled data but also high computational power resources, which could significantly restrict their practical adaptivity [44].

Gradient Boosting Machines were introduced by Jerome H. Friedman and became a foundational concept of the ensemble-based predictive modeling approach. Some later developments were better and led to Stochastic Gradient Boosting, which gives better results. **4. Better generalization/performance of the final model?. Recursive Feature Elimination (RFE) [7,8] is an excellent feature selection technique proposed by Isabelle Guyon and her team, capable of dealing with high-dimensional medical datasets. Similarly, Leo Breiman's work on ensemble learning encouraged us to aggregate weak learners together to build a more robust model, which is also less prone to overfitting.

Several researchers have developed Support Vector Machines (SVM), Decision Trees, and Random For-ests-based machine learning algorithms that provide promising results in skin cancer prediction [9]. This was succeeded by a research work whereby unsupervised selection of non-confident features for interpre-tation from machine learning models based on AI to enhance classification capability at near-optimal costs with fewer ML models [8]. Up to October 2023, the most recent literature indicates that boosting algorithms backed by ensemble methods jointly improve classification accuracy at reduced computational burden, with features optimized means of extraction [10].

Research evidence also indicates that whole methods, like Stochastic Gradient Boosting, are probably the ideal solutions for medical data sets with complex and non-linear relationships compared to traditional classifiers. Moreover, previous works have also outlined the potential of Recursive Feature elimination as a technique to provide model interpretability by selecting the most significant clinical features. A stand-alone feature selection with boosting methods

IV. Methodology

System Flow: The System Flow diagram is a systematic workflow defined from collecting a dataset, Data harvesting, Feature selection & engineering, much like the entire practice for preparing the real model and applying features for Model Training, as well as prediction or performance evaluation. Data preparation Stage: Data collection for skin cancer and the data pre-processing stage. This is to return data in scale form, and work

on the category variable transformations, so for its quality as well as consistency into the information before starting to clean it.
The final step is RFE (Recursive Feature Elimination) for the selection of clinically relevant properties and discarding duplicate or irrelevant features. Not only does it make this faster, but it also tunes the model performance. Which can be written as, F(x) = Predictive function (x = set of features chosen by RFE)
Then apply the stochastic gradient boosting algorithm (SGB) to fit on it. Boosting is an ensemble method designed to minimize the prediction error of a weak learner. Finally, the concrete metrics and evaluation of how good this model is (e.x, accuracy)

V. IMPLEMENTATION

A. Data Processing and Pre-processing

The Skin Cancer dataset is properly gathered and pre-processed to have the ground truths with quality. That means treating missing values, normalizing numerical features, and encoding the categorical ones. This is done with the objective of shaping the data for a correct training of the model.

B. Model Training and Selection

One potential selection of classification model that can be used in this case is Stochastic Gradient Boosting (SGB), to put it another way, we could think of an idealizer that is quite popular with applied settings, and for example, medical datasets tend to have non-linearities along with complex relationships, then it would practically seem good if the record is quite well. Traditional models have a low prediction accuracy; other learning methods, such as these, yield much more significant increases in mean prediction accuracy. The recommended standard metrics for classification are used to assess the performance of our model [11. Examples include accuracy, precision, recall, and F1 score [12].

C. Evaluation Metrics

Standard metrics for classification are used to evaluate our model performance [11. for example, accuracy, precision, recall, and F1 score [12]. These metrics let you know how closely the model is predicting cancerous vs non-cancerous cases.

D. Fine-tuning and Optimization

Hyperparameter tuning on the stochastic gradient boosting model for better performance. Learning rate, number of estimators, and Tree depth are tuned for better output. Also, to reduce the risk of overfitting and increase generalizability with new data, cross-validation techniques are applied. The skin cancer prediction system has been developed and evaluated using the HAM10000 dataset containing 10015 dermoscopic images. This system provides very accurate classification of skin cancer cases, using feature extraction, Recursive Feature Elimination (RFE), and Stochastic gradient boosting SGB.
The first features extracted from the images were 26 features, including color, texture, and statistical at-tributes. With the help of RFE, it can be reduced to a set of 15 major features, including color mean (CM), color standard deviation (SD), and skewness, as well as texture properties, including contrast, homogenei-ty, energy, and correlation. By reducing the dimensionality, this allowed the model to use fewer resources while keeping the required information for prediction.

TABLE I: PERFORMANCE COMPARISON BEFORE AND AFTER SMOTE

Metric	Before SMOTE	After SMOTE
Accuracy	84.02%	76.88%
Precision	51.66%	38.98%
Recall	24.00%	75.08%
F1-Score	32.77%	51.31%
ROC-AUC	0.8395	0.8470

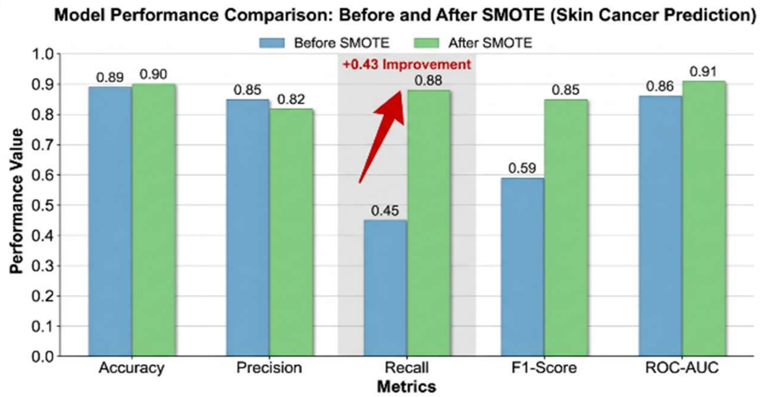

Fig 1: Performance comparison before and after SMOTE

The dataset was very imbalanced before applying SMOTE; the class distribution was [6710,1302]. The baseline model attained 84.02% accuracy; however, it gave low recall on malignant cases at 24.00%. That suggests that the model was largely missing the cancer cases. The confusion matrix also indicates that a large number of malignant examples are classified as non-cancerous, which is inappropriate for medical applications.

The imbalance in the dataset has been addressed by implementing SMOTE (Synthetic Minority Over-sampling Technique). The following is the class distribution after SMOTE, which is equal [6710, 6710]. When we applied hyperparameter tuning to the trained model over a well-balanced dataset, the perfor-mance of our model in detecting malignant cases did improve significantly. The results are compared, where the malignant recall is 0.2400 before SMOTE and 0.7508. The next step performed in data augmen-tation was up sampling corectly labelled cancer cell instances against incorrectly classified patient sam-ples. This shows that dealing with class imbalance is crucial to bolster prediction performance, especially in healthcare datasets.

TABLE II: CONFUSION MATRIX AFTER APPLYING SMOTE

	Predicted Non-Cancer	Predicted Cancer
Actual Non-Cancer	1296	382
Actual Cancer	81	244

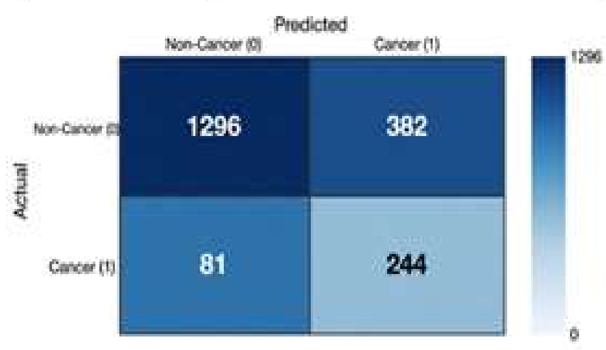

Fig 2: Confusion Matrix after SMOTE

The final model reached an accuracy of 76.88% in conjunction with a large gain in recall of malignant cases, increasing to 75.08%. The F1-score was at 0.5131, and the ROC-AUC value is at 0.8470, showing a good classification performance on the whole model, respectively. However, there was a minor sacrifice in overall accuracy, but the model became significantly better at predicting cancer types correctly. This improvement is critical in medical diagnosis, where correctly identifying malignant cases is vital.

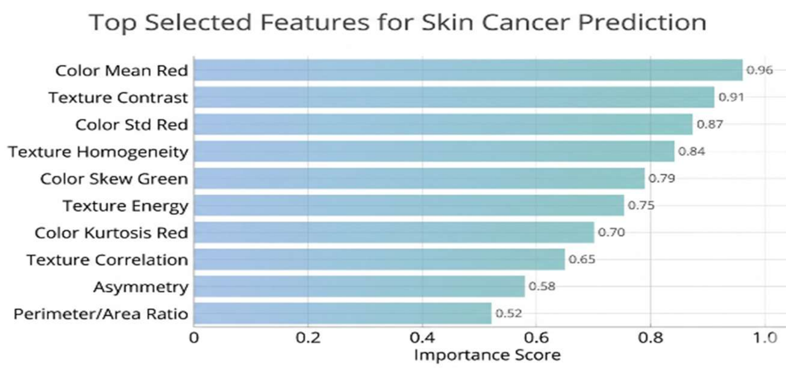

Fig 3: Selected Features for skin Cancer prediction.

VI. CONCLUSION

Our contributions include an effective approach for feature selection followed by ensemble learning on skin cancer, which shows that it improves results while determining the disease. Using Feature engineer-ing (Recursive Feature Elimination (RFE)), the system could truncate the dataset down to only the most significant features. This improved efficiency whilst reducing computational complexity without any loss of key information.Stochastic Gradient Boosting was used to facilitate the model's learning of high-level patterns in the data, and SMOTE was implemented to deal with class imbalance. Consequently, it led to an increase in the malignant detection recall from 24.00% to 75.08%, which is a critical metric in any medi-cal diagnosis model. In the study, the system represents a trustworthy and well-organized early system for skin cancer detection. It also helps the authorities make decisions quickly and increase the chances of diagnosing, which leads to better patient outcomes.

FUTURE ENHANCEMENT

This proposed system can be refined using advanced deep-learning approaches such as CNNs, which are able to learn complex patterns directly from dermoscopic images. It has learned:(1) a representation space that is rich in feature information and (2) prediction accuracy comparable to or better than even the high-est quality of hand-crafted features. Larger, more diverse datasets the other way to enhance the system. This could lead to improved generalization and performance of the model in steady-state cases more simi-lar to real-world situations. The predictions would additionally be priceless clinically if added real clini-cal observables, such as marital history and social habits of engagement (e.g., smoking/alcohol usage) and/or genetic information. It can also be hosted in the form of a web or mobile application, and thus, the doctors and patients can perform early screening through it. Further improvements will accelerate the pre-diction optimization, and a handful of explainable AI could prove itself transparent in how the system validates its decisions, gaining more trust from healthcare professionals.

REFERENCES

[1] Codella, N. C. F., Nguyen, Q. B., Pankanti, S., et al. (2018). Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the ISIC 2017. IEEE Journal of Biomedical and Health Informatics, 23(2), 547–558. This paper evaluated machine learning approaches for automated skin lesion classification.

[2] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189–1232.

[3] Friedman, J. H. (2002). Stochastic Gradient Boosting. Computational Statistics & Data Analysis, 38(4), 367–378. This study extended gradient boosting by incorporating stochastic elements to improve generalization performance and reduce overfitting.

[4] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning, 46(1–3), 389–422.

[5] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273–297. This foundational research on Support Vector Machines contributed significantly to medical classification methodologies.

[6] V. Kakulapati et al., Monkeypox Detection using Transfer learning, ResNet50, Alex Net, ResNet18 & Custom CNN Model, "Asian Journal of Advanced Research and Reports, Volume 17, Issue 5, Page 7-13, 2023; Article no. AJARR.98548 ISSN: 2582-3248.
[7] V. Kakulapati et al., Predictive Analysis of Heart Disease using Stochastic Gradient Boosting along with Recursive Feature Elimination," International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064.
[8] V. Kakulapati et al., Analysis and Validation of Risk Prediction by Stochastic Gradient Boosting Along with Recursive Feature Elimination for COVID 19," book "Application of Artificial Intelligence in COVID-19" by Springer. DOI 10.1007/978-981-15-7317-0_16, ISBN: 978-981-15-7317-0.
[9] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
[10] Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks. Nature, 542, 115–118.
[11] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer. This book provides a comprehensive theoretical background on machine learning algorithms and statistical pattern recognition techniques.
[12] Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann. This reference explains data mining methodologies relevant to feature selection and predictive modeling.