Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 252 (2025) 317–330

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Evolutionary Computing and Mobile Sustainable Networks

Classification of Sarcoma Based on Genomic Data Using Machine Learning Models

Pratham Gala[a], Yash Pandloskar[a], Shubham Godbole[b], Fayed Hakim[b], Pratik Kanani[b], Lakshmi Kurup[b]

[a]*Dept. of Computer Science and Engineering (Data Science), Dwarkadas J. Sanghvi College of Engineering, Mumbai, India*
[b]*Dept. of Artificial Intelligence and Data Science, Dwarkadas J. Sanghvi College of Engineering, Mumbai, Indi*

Abstract

The proposed work provides a new machine-learnt classification approach for the various types of soft tissue sarcoma based on genomics data which addresses a considerable gap in sarcoma diagnostics. The previous studies have investigated various aspects of sarcoma but this study is unique in that it targets the predicting sarcoma variant types using genetic information, which has not been done before. Random Forest was used as the meta-estimator and a stacking ensemble model comprising of Random Forest, Extreme Gradient Boosting and LightGBM were used for this study. The model which was trained and validated on a complete dataset of 206 adult soft tissue sarcoma samples containing genomic alterations, transcriptomic, epigenomic and proteomic data achieved an accuracy of 89.44% at a precision level as high as 91%. Stratified k-fold cross validation is employed to ensure that class imbalance is not a hindrance to performance. This innovative approach outmatches single classifiers and traditional single model methods at great length hence making it possible and effective to use machine learning on genomic data for predicting sarcoma variants. Thus, the findings from this research could change cancer diagnosis forever; they promise more accurate classification as well as personalized treatment modalities while also providing a framework for analogous applications in other rare complex cancers.

Keywords: Sarcoma Classification; Stacking Ensemble Model; Genomic Data

1. Introduction

Sarcomas are a broad category of cancerous tumors originating from connective tissues, including cartilage, bone, muscle, and fat. Sarcomas are extremely rare, making up about 1% of adults and 15% of pediatric malignancies. However, due to their aggressiveness and multifariousness, they pose a major problem in clinical oncology [1].

E-mail address: galapratham03@gmail.com

1877-0509
10.1016/j.procs.2024.12.034

Histopathological examination is a key component of the traditional classification of sarcomas, nevertheless, it can be subjective and highly prone to mistakes. Therefore, there is an exigency for more accurate and objective approaches to categorize such tumors to make treatment choices effectively and enhance patient outcomes.

Recent advancements in genomic technology have produced an abundance of data, presenting fresh possibilities for sarcoma molecular characterisation. Comprehension of the biological diversity and underlying mechanisms of these cancers may be improved by the integration of genomic data, such as DNA sequencing, RNA expression patterns, and epigenetic markers. However, analysis and interpretation of genomic data are severely hampered by its enormous dimensionality and complexity [2].

Large-scale genomic datasets may be handled by machine learning (ML) models, which can also be used to find patterns that conventional statistical techniques might miss. These models can be taught to identify complex patterns in the data and forecast outcomes based on those patterns. ML models are a viable method for creating more precise and complex classification systems that take genetic information into account when it comes to sarcoma.

The purpose of this work is to investigate how machine learning methods can be applied to the genomic classification of sarcoma. In order to categorize sarcoma subtypes based on genomic profiles, this paper considers a variety of machine learning (ML) algorithms, including supervised learning techniques like support vector machines (SVM), random forests, and deep learning networks. This work also analyzes the accuracy, sensitivity, and specificity of these models' performance and determines whether they have the ability to improve on the state-of-the-art diagnostic procedures [3].

The following study aims to further the creation of more accurate categorization frameworks for sarcoma by utilizing machine learning techniques, which could ultimately result in more individualized treatment plans and more accurate prognosis assessments [4-6]. The proposed work aims to exhibit how ML models may be incorporated into clinical workflows and demonstrate the promise of precision medicine powered by genomics in the treatment of sarcoma.

2. Literature Survey

The literature survey encompasses three key areas relevant to the proposed research: soft tissue tumors, machine learning applications in medical diagnostics, and genomic classification techniques. Soft tissue tumors, due to their diverse nature, present significant challenges in diagnosis and prognosis, necessitating highly effective diagnostic methods. The application of machine learning algorithms in medical sciences has demonstrated potential for improving the accuracy and efficiency of tumor classification and prediction. Furthermore, genomic classification approaches have shown promise in enhance the understanding of the biological basis for histological diagnoses, potentially leading to improved therapeutic options for patients. This survey aims to provide a comprehensive overview of these interconnected fields, highlighting current advancements, methodologies, and potential areas for further research. Despite the excellent contributions in these areas, there remain aspects that require additional investigation, particularly in the integration of machine learning techniques with genomic data for more precise sarcoma classification.

2.1. Soft Tissue tumors

Abnormal growths in the body's soft tissues, such as muscles, fat, tendons, blood vessels, and nerves, are known as soft tissue tumors. These growths may be malignant (cancerous) or benign (non-cancerous). While benign soft tissue tumors are more common and usually do not spread to other parts of the body, malignant tumors, also known as soft tissue sarcomas, can be aggressive and have the potential to spread to other organs. The kind, size, location, and malignancy status of soft tissue tumors often have an impact on diagnosis and treatment.

David van IJzendoorn et al. [7] focuses on soft tissue sarcomas and analyzes gene expression data using machine learning techniques in an effort to find new diagnostic and prognostic markers as well as potential therapy targets. This is accomplished by combining data from The Cancer Genome Atlas (TCGA) with the Genotype-Tissue Expression (GTEx) project. By applying a deep neural network technique to identify similarities between tumor samples and normal tissues, the research attains a high prediction accuracy.

A dataset comprising of information on several medical tests, such as complete blood counts (CBC) and blood correlation, and that comprises 50 patients who had soft tissue tumors (STT) diagnosed and 25 patients who had STT misdiagnosed was employed by Abdellaoui Alaoui et al. [8]. By developing and assessing machine learning models that accurately differentiate between STT and non-STT cases (based on patient data analysis), the research seeks to improve automated STT diagnosis.

A research focused on the understanding of molecular and genetic changes occurring in a few selected soft tissue tumors including gastrointestinal stromal tumors (GISTs), peripheral nerve sheath tumors and epithelioid vascular tumors was conducted by Schaefer IM et al. [9]. The study sets out to enhance the classification and management of soft tissue sarcomas by including these molecular insights to clinical practice since there is a paradigm shift in the practice of oncology.

2.2. Machine Learning Algorithms

In the field of medicine, instead of being referred to as managers only, machine learning models are built and deployed in a novel manner to take on diagnosis, prognosis, and therapy planning. This is so because algorithms are formed to assess complex medical information like that of patient history, genetic factors, and clinical images. Such algorithms can successfully identify the patterns and insights that may seem enigmatic to a human, making it feasible to detect the disease at the earliest stage, design the most effective treatment plans personalized to the individual, and consequently save more lives. Machine learning models are evolving and have the potential to disrupt many areas in medicine.

Jaber Juntu et al. [10] performed tumor classification against soft tissue area targets employing machine learning methods and texture measures derived from MR images. A variety of different features are incorporated into machine learning classifiers such as decision trees, neural networks and support vector machines which are then trained and tested to see how powerful they are using such features. The analysis examines the effect of size and quality of the training dataset on the performance of the models with adequate statistical tests and cross-validations to avoid overfitting the classifiers.

Developments are made in the area of constructing a consistent classification model that uses gene expression data to distinguish various types of cancer, as demonstrated by Jun Xiao et al. [11] in their study. This employs a dataset of 801 samples of lung adenocarcinoma (LUAD), clear cell kidney carcinoma (KIRC), prostate adenocarcinoma (PRAD) and colon adenocarcinoma (COAD), five cancers types. The approach makes use of exploratory data analysis, dimensionality reduction and various clustering along machine learning techniques such Decision Trees, SVM, Random Forest, Naive Bayes, and Deep Neural Network.

A Stacking Classifier model was developed to improve the accuracy of diabetes detection by utilizing various machine learning techniques, as shown by Maria Ali et al. [12]. The research splits the dataset into training and testing subsets using K-Nearest Neighbors (KNN), Naive Bayes, Linear Discriminant Analysis (LDA), and Decision Tree as basic classifiers to generate predictions. These predictions are then used to create a new dataset, which is fed into a Random Forest meta classifier, in an effort to increase the accuracy of the final prediction.

According to Alexandropoulos SA, N. et al. [13], for increasing classification accuracy, several classifiers can be stacked together in a hybrid ensemble approach. In order to construct a robust meta-classifier that utilizes the benefits of individual base learners, this work combines several ensemble techniques such as Random Forest, Gradient Boosting, and ExtraTrees. The process of stacking combines base classifiers' predictions and makes use of these in another learning algorithm to produce final predictions.

This approach has led to improvement of the classifier fusion systems in the form of stacking as articulated by Dzeroski S. et al. [14]. The authors assess the effectiveness of existing stacking techniques and propose improvements employing MLR, the meta-level learning technique, and a broader array of meta-level features. The aim is to support through experimentation with several datasets that such extensions can be more effective for achieving improved classification as compared to the existing methods thus advancing the field of meta-learning and strategies for combining classifiers.

2.3. Genomic Classification

Machine learning designed genomic classification refers to an advanced technique that involves using computational technologies in analyzing and interpreting genetic material. Labeled as genetics-based pattern recognition systems, it involves the construction and training of models that are fed with genomic big data to recognize genetic elements such as genes, mutations, and other variants contained within DNA sequences. This technique is crucial in understanding complex biological systems, identifying disease-causing mutations, and providing personalized therapies. Genomic categorization is an emerging strategy in genomics and precision medicine because of potential improvement in the disease prediction accuracy, congenial medicines, and understanding the genetic basis of disease heterogeneity.

Taylor BS et al. [15] proved focused on the analysis of the molecular mechanisms of prostate cancer via genetic alterations in cancerous tissue and finding treatment options for the disease. The study finds certain genetic alterations that are frequent occurrences in prostate cancer like the TMPRSS2-ERG gene fusion and other chromosomal changes through large-scale genomic approaches. An intended impact of the findings is to improve the quality of treatment meted out to patients with prostate cancer by optimizing the classification of patients according to genetic risk factors and defining better the molecular characteristics of prostate cancer.

As shown in, The Cancer Genome Atlas Research Network examines the genetic features of the tumors using a number of models and techniques. Important models include the integrated clustering model called iCluster, which integrates SCNAs, DNA methylation, mRNA and microRNA expression data, and facilitates the discrimination of squamous cell carcinoma from adenocarcinomas [16].

2.4. Research Gaps

Based on the comprehensive literature survey encompassing soft tissue tumors, machine learning algorithms, and genomic classification techniques, several key research gaps have been identified. These gaps represent areas where current approaches fall short or where there is significant potential for improvement in the field of sarcoma classification using genomic data and machine learning. By analyzing the methodologies, results, and limitations discussed in the reviewed studies, the following gaps have been synthesized that warrant further investigation and development:

1. Challenges in managing high-dimensional genomic data and selecting relevant features consistently.
2. Risk of models performing well on training data but poorly on unseen data, especially with limited sample sizes.
3. Inconsistent model building procedures and evaluation metrics across studies, hindering reproducibility.
4. Many studies focus on single data types, missing the potential insights from integrating multiple genomic and clinical data sources.

3. Methodology

The process of developing an optimal model for sarcoma classification involved extensive experimentation and iterative refinement. Here's a detailed account of the process:

3.1. Work-Flow

The given diagram is the workflow was followed for the proposed architecture: Fig 1, describes the flowchart for the entire process, the first column represents the main steps followed in the procedure, the subsequent row describes the subdivisions of the tasks done to fulfill the main step.

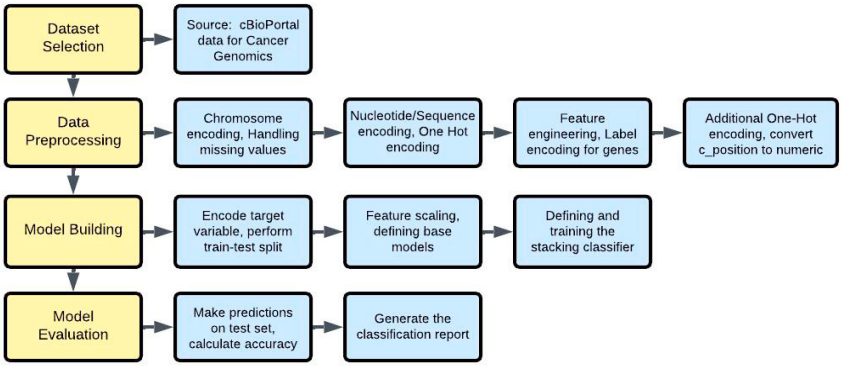

Fig. 1: Flowchart of Machine Learning-based system

- Data Splitting: The preprocessed data is split into training (80%) and testing (20%) sets, maintaining the distribution of sarcoma types.
- Base Classifier Training: Each base classifier (XGBoost, LightGBM, Random Forest) is trained on the training data. 5-fold cross-validation is used to generate out-of-fold predictions.
- Feature Creation for Meta-Classifier: Predictions from base classifiers on the validation folds are used as features for the meta-classifier.
- Meta-Classifier Training: The Random Forest meta-classifier is trained on the new feature set created from base classifier predictions.
- Hyperparameter Tuning: Grid search with cross-validation is employed to optimize hyperparameters for both base and meta-classifiers.
- Final Prediction: The trained stacking model (an ensemble of XGBoost, LightGBM, Random Forest) makes predictions on the test set.
- Model Interpretation: Feature importance analysis is conducted to identify key molecular drivers of sarcoma subtypes and outcomes.
- Performance Evaluation: Performance metrics including accuracy, precision and recall are evaluated. Performance is compared against individual base classifiers and traditional single-model approaches.

To improve the accuracy and precision of the classification models, data cleaning, integration, normalization and data wrangling can be done as a part of the pre-processing steps. Secondly, ensemble learning methods where several base models are combined into one stacking classifier can be done to improve the accuracy. Another few techniques like undersampling and oversampling along with optimizing the model using grid search techniques are used to achieve desired targets.

3.2. Dataset Description

The present study employs new approaches to gather data using very complex upon analysis adult soft tissue sarcomas data set available in cBioPortal database. This dataset, [17], is made up of 206 adult SOFT tissues sarcomas samples and represents six major types thus giving a wide optic view of the genomic samples of these mesenchyinal tumors. It comprises of:

- Genomic alterations: Copy number variations, mutations, or structural variants.
- Transcriptomic data: RNA-Seq.
- Epigenomic data: DNA methylation.
- Proteomic data: Protein expression.
- Clinical data: Outcome of patients in terms of their clinical features.

This rich dataset allows for an integrated exploration of the molecular features associated with sarcomas, which could help understand the heterogeneity of these tumors and identify new opportunities for treatment.

3.3. Data Preprocessing

To reduce the imbalance and clean the data, the following steps were employed:

- Data Cleaning and Integration:
  - Data from various sources – genomic, transcriptomic, epigenomic, proteomic information were combined in merged file
  - Deleted samples with too much missing data.
  - Cross type standardization of gene identifiers.
  - Dropped columns 'Expression_value', 'Hugo_Symbol', 'Entrez_Gene_Id', 'dbSNP_RS', 'dbSNP_Val_Status', 'Verification_Status', 'Validation_Status', 'Mutation_Status, 'Sequencing_Phase', 'Sequence_Source', 'Validation_Method', 'Score', 'BAM_File'
- Feature Engineering:
  - Genomic Data : Chromosome numbers were converted to number representation.
  - Mutation Data : It was necessary to encode positions of coding sequences. Created binary features for mutation presence in key genes (e.g., TP53, ATRX, RB1).
  - Copy Number Data : Segmented copy number data into discrete categories.
  - Expression Data : Normalized expression values. Calculated pathway activation scores based on gene set enrichment analysis.
  - Methylation Data : Computed beta values for CpG sites. Aggregated methylation data at the gene and promoter levels.
- Feature Selection:
  - Utilized domain knowledge to select relevant features.
  - Employed correlation analysis and mutual information to identify informative features.
- Encoding:
  - Applied one-hot encoding to categorical variables such as sarcoma types, validation status.
  - Used label encoding for gene symbols to create columns and then removing the columns via feature engineering.
  - Labeled chromosomes X, Y and MT as 100, 101 and 200 respectively
- Handling Class Imbalance:

  'Missense mutations' had a high majority due to which it was undersampled by dropping values to reduce bias. 'Silent' and 'missense mutations' were a minority for which were oversampled by replication to ensure adequate representation.
- Data Normalization:

Standard scaling is used to normalize and reduce size of numerical features to ensure comparability.

- Missing Data Imputation:

  K-Nearest Neighbours (k-NN) was implemented to impute missing values for features with partial missing data.

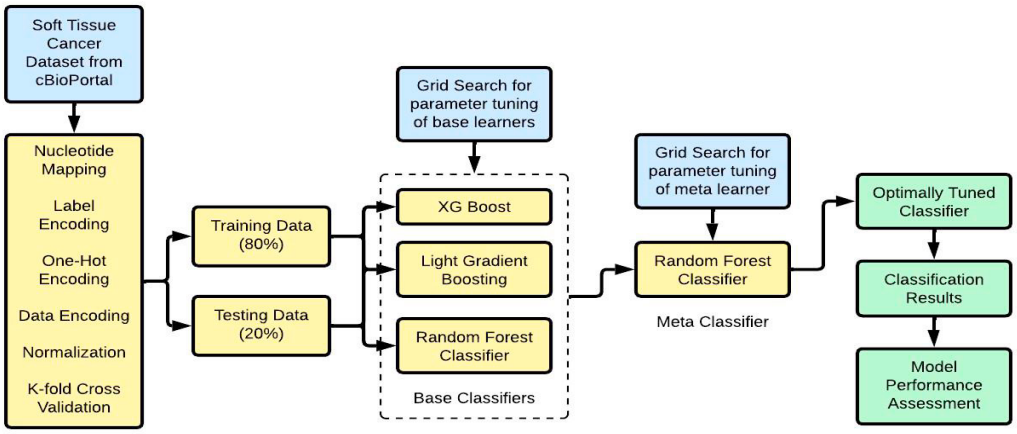

Fig. 2: Proposed Model: Stacking Classifier

4. Proposed Model

The proposed approach to sarcoma classification utilize the power of ensemble learning, specifically employing a stacking classifier architecture. This advanced technique combines multiple base models with a meta-classifier to create a robust and high-performing predictive system. The stacking classifier was chosen after extensive experimentation with various individual models and ensemble methods, as it demonstrated superior performance in handling the complexity and high dimensionality of genomic data. The architecture consists of three diverse base classifiers - XGBoost, LightGBM, and Random Forest - each bringing unique strengths to the ensemble. These base models were tuned by training on multiple sets of hyperparameter using GridSearcCV and the best performing parameters were chosen. These base models then feed their predictions into a Random Forest meta-classifier, which learns to optimally combine their outputs for final prediction. This approach allows for capturing of complex patterns in the genomic data that might be missed by single models, leading to improved accuracy and generalization in sarcoma subtype classification.

- Base Classifiers:

  (a) XGBoost: • Handles non-linear relationships in genomic data.
  - Effective in capturing complex interactions between features.

  (b) LightGBM:
  - Efficient for high-dimensional data.
  - Captures fine-grained patterns in molecular profiles.

  (c) Random Forest: • Robust against overfitting.
  - Effectively handles the mix of categorical and numerical features in the dataset.

- Meta-Classifier:

(a) Random Forest: Synthesizes predictions from base classifiers.

Let X be the input feature vector and y the true label.

For base classifiers f_1, f_2, and f_3:
$\hat{y}_1 = f_1(X), \quad \hat{y}_2 = f_2(X), \quad \hat{y}_3 = f_3(X)$
The meta-classifier g combines these predictions:
$\hat{y}_{\text{final}} = g(\hat{y}_1, \hat{y}_2, \hat{y}_3)$ (1)
Where:
f_1 : XGBoost classifier, f_2 : LightGBM classifier, f_3 : Random Forest classifier, g : Random Forest meta-classifier
The stacking process minimizes the loss function:
$\mathcal{L} = \sum_{i=1}^{n} L(y_i, g(f_1(X_i), f_2(X_i), f_3(X_i)))$ (2)

5. Result Evaluation and Discussion

In order to guarantee the correctness and dependability of research findings, extensive testing is necessary for validation. Test cases are essential to this procedure because they offer a methodical way to assess different facets of the study. Test cases help identify any problems by providing precise scenarios and expected results. This guarantees that the research can survive criticism and make a significant contribution to the subject.

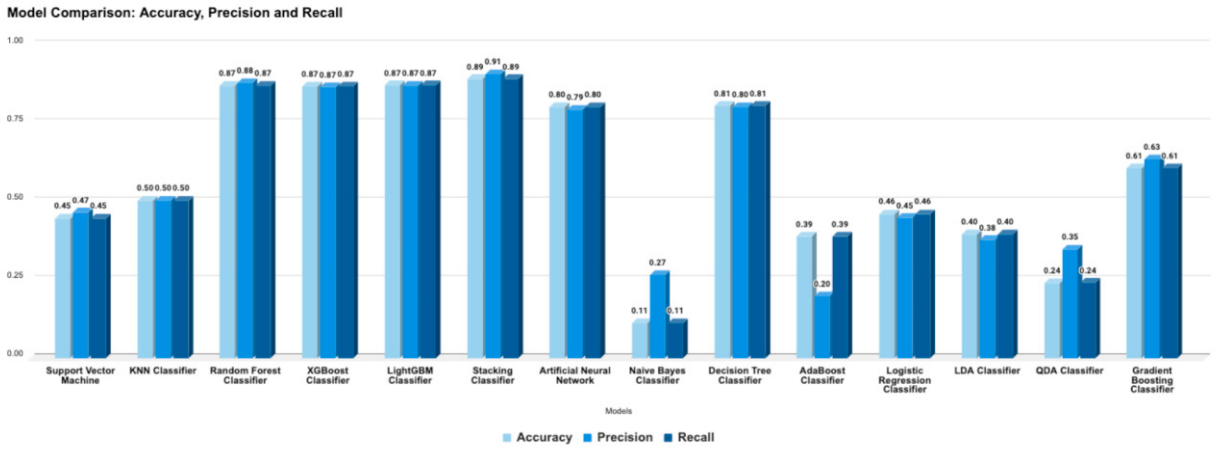

Fig. 3: Comparison of Effectiveness of Different Models

In fig 3, the bar chart shows results of performance evaluation metrics of the several models such as accuracy, precision and recall for Support Vector Machine, KNN, Random forest, Xgboost, Light GBM among other models. The 'stacking classifier' scores on top in terms of metric values such as accuracy, precision, recall, more than 89 to 91% and all other SA class struggles are on comparison in lower models like Naive Bayes, AdaBoost, and QDA is lower than the 50% mark in other metrics performance scores as its low within the area under the curve. It appears that configuration of the models, namely Stacking; Random Forest, LightGBM, are very useful in this case because of the nature of the data present, which is genetic data which is highly complex in the general approaches of prediction. Table 1 describes the metrics the models were compared against. As shown in table 1, tree based classifiers produced the best results followed by Artificial Neural Networks. These models had noteworthy results while the rest were subpar. Table 2 consists of mutation types and their corresponding index for the indices used in confusion matrices. Another metric that can be considered is the classification error, which is the proportion of incorrect predictions, which can be calculated by taking compliment of accuracy.

$$\text{Classification Error} = \frac{\text{Number of Incorrect Predictions}}{\text{Total Number of Predictions}} \tag{3}$$

Table 1: Comparison of Model Metrics

Models	**Accuracy**	**Precision**	**Recall**
Support Vector Machine	0.4468	0.4655	0.4468
KNN Classifier	0.5039	0.5041	0.5039
Random Forest Classifier	0.8704	0.8784	0.8704
XGBoost Classifier	0.8677	0.8661	0.8677
LightGBM Classifier	0.8715	0.8694	0.8715
Stacking Classifier	**0.8914**	**0.9072**	**0.8915**
Artificial Neural Network	0.8021	0.7938	0.8021
Naive Bayes Classifier	0.1131	0.2654	0.1131
Decision Tree Classifier	0.8063	0.8021	0.8062
AdaBoost Classifier	0.3885	0.1981	0.3885
Logistic Regression Classifier	0.4602	0.4479	0.4602
LDA Classifier	0.3961	0.3771	0.3961
QDA Classifier	0.2407	0.3473	0.24076
Gradient Boosting Classifier	0.6069	0.6346	0.6069

Table 2: Legend of Mutation Types for Confusion Matrices

No.	Mutation Type
0	Frameshift Mutations
1	Inframe Mutation
2	Missense Mutation
3	Nonsense Mutation
4	Other Mutation
5	Regulatory Mutation
6	Silent
7	Splice Variant

The results of XG Boost is shown in fig 4, it depicts the per class classification of the model. It produced favorable results which acted as a baseline for the other tree-based classifiers.

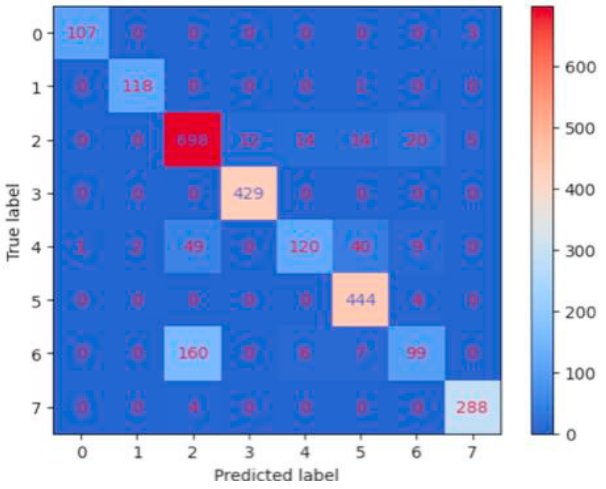

Fig. 4. Confusion Matrix for XG Boost

In fig 4, it is observed that excellent performance is observed with XG Boost in classifying Frameshift Mutations (107/110 correct), Inframe Mutations (118/119 correct), and Variants (28 Splice8/292 correct). Strong performance for Non-sense Mutations (429/429 correct) and Regulatory Mutations (444/448 correct). Some misclassification of Missense Mutations as Silent (160 cases) and Other Mutations (49 cases).

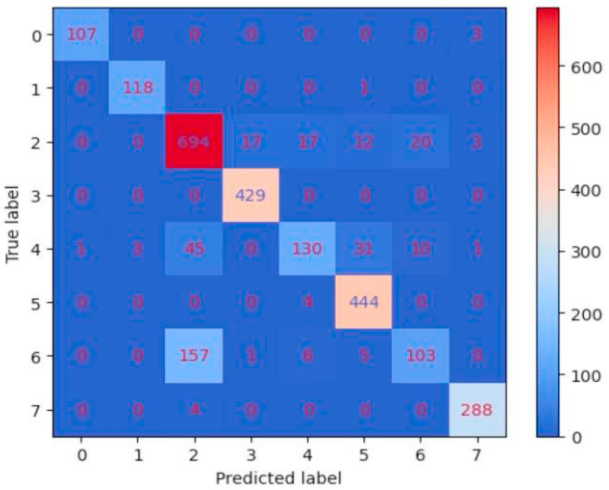

Fig. 5: Confusion Matrix for Light Gradient-Boosting Machine

As shown in fig 4 and fig 5, a similar pattern to XGBoost is observed in LightGBM, with slightly better performance on Missense Mutations (715 correct vs 698 for XGBoost). Slight improvement in classifying Silent mutations (91 correct vs 99 for XGBoost). Maintains high accuracy for Frameshift, Inframe, Nonsense, and Splice Variants.

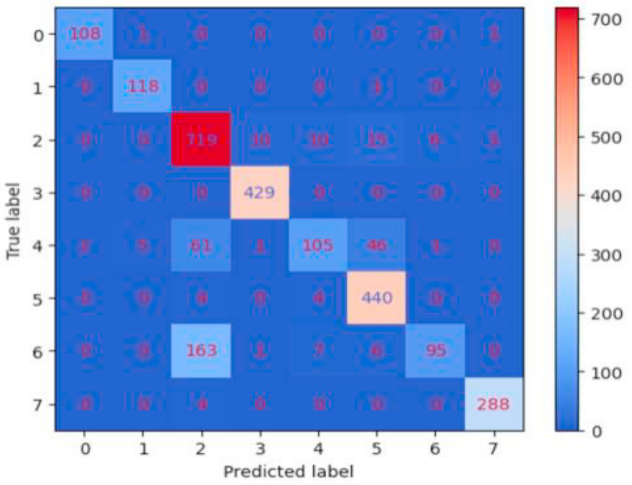

Fig. 6: Confusion Matrix for Random Forest

As understood from fig 6, the performance is very similar to XG Boost in Random Forest, with identical or near-identical results for most mutation types. Slightly fewer misclassifications of Missense Mutations as Silent (160 vs 167 LightGBM).

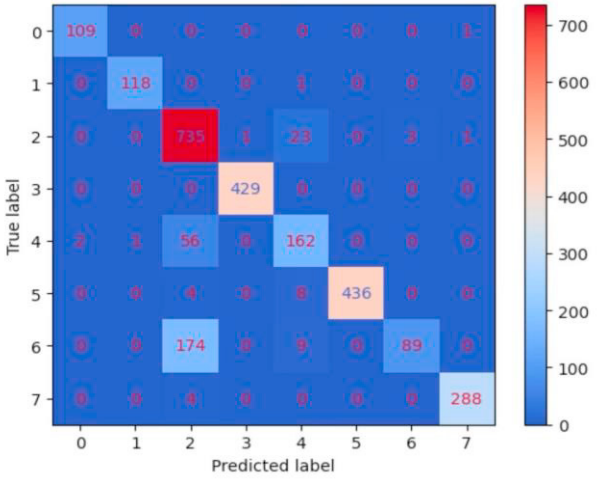

Fig. 7: Confusion Matrix for Stacking Classifier

In fig 7, the results of the Stacking Classifier are identical to that of the Random Forest Classifier. This suggests that the stacking approach, while improving overall metrics, maintains the same pattern of strengths and weaknesses as the Random Forest model in terms of class-specific predictions.

The proposed stacking classifier performed exceptionally well, attaining an accuracy rate of 89.14%, demonstrating the effectiveness of ensemble learning in managing complex genomic data. This method successfully combines the advantages of many base classifiers (XGBoost, LightGBM, and Random Forest) to identify a variety of patterns in the data that a single model could overlook. Despite the dataset's class imbalance, the confusion matrices show that The proposed model performs exceptionally well in categorizing uncommon mutation types like Frameshift, Inframe, and Splice Variants. This shows that the model is capable of recognizing unique genomic signatures linked to these less common sarcoma subtypes, which is important for enhancing the detection of less common variations. The fact

that the proposed model consistently performs well on several evaluation measures (recall, accuracy, and precision) shows that it is not only accurate but also evenly distributed in its predictions among various sarcoma subtypes. This is especially crucial when it comes to patient care in a clinical setting, as false positives and negatives can both have serious consequences.

Table 3: Execution time for proposed models

Model	Time Taken for 1000 predictions (s)
Random Forest Classifier	0.033
XGBoost Classifier	0.199
LGB Classifier	0.047
Stacking Classifier	0.283

As shown by the table above Random Forest is the fastest in terms of making 1000 predictions in the span of 0.033s. While Stacking Classifier is the slowest at 0.283 s for 1000 predictions. The tradeoff for a quicker prediction is lower precision, as discussed before Stacking Classifier has a better precision and class-wise accuracy.

Key Observations

1. The dataset shows a significant imbalance, with Missense Mutations being the most common class. This imbalance is well-handled by all models, but particularly by the Stacking Classifier.
2. All models show excellent performance in classifying rarer mutation types like Frameshift, Inframe, and Splice Variants, indicating robust learning despite class imbalance.
3. The main area of misclassification across all models is Missense Mutations, often confused with Silent mutations. This suggests a need for further feature engineering or more advanced techniques to distinguish these closely related classes.
4. While the Stacking Classifier's confusion matrix is identical to the Random Forest's, its superior overall metrics (as seen in the bar graph) suggests it is making more confident and accurate predictions across all classes.
5. The same model is also tested against sarcoma test cases from [18]. The produced results have generated q.

6. Conclusion

This work demonstrates the effectiveness of a machine learning approach in classifying soft tissue sarcoma subtypes using genomic data. A range of ML models were evaluated, including SVM, KNN Classifier, Random Forest, XGBoost Classifier, LightGBM Classifier, ANN, Naive Bayes, Decision Tree, AdaBoost Classifier, Logistic Regression, LDA Classifier, QDA Classifier, Gradient Boosting Classifier, and a stacking classifier combining XGBoost, LightGBM, and Random Forest. Tree based classifiers; Random Forest, XGBoost Classifier, LightGBM Classifier, Decision Tree and AdaBoost Classifier achieved higher accuracy and produced better results than other models. XGBoost, Random Forest and LightGBM achieved an accuracy of 86.77%, 87.04%, 87.15% respectively were used to develop a stacking classifier which achieved an accuracy of 89.14%. The stacking classifier yielded the highest overall accuracy and minimized misclassifications more effectively than the individual models, it also provides multiple classifications in the span of seconds as discussed in table 3 which would allow for multiple diagnoses. The significance of this research lies in its potential to revolutionize sarcoma diagnosis and treatment. By accurately classifying sarcoma subtypes based on genomic data, the proposed model addresses a critical need in precision oncology. This approach can enhance diagnostic accuracy, accelerate research by facilitating the identification of new biomarkers and potential therapeutic targets for sarcomas. The success of this architecture advocated the notion that computational methods are very effective in revealing hidden patterns in high dimensional genomic datasets that cannot be seen by conventional ways of analysis.

7. Future Scope

The proposed work shows the effectiveness of ML models in classifying subtypes of sarcomas. Future directions for further research include external validation, where the models generalizability is tested against independent datasets from various clinical settings. Another path lies in refining the approach proposed in this work to better identify and classify rare sarcoma subtypes. Further research in Pan-Cancer applications can help develop a unified solution for cancer classification based on genomic profiles.

In order to ensure reliability, the study may adopt cross-validation procedures, use an independent data set for external validation, and perform sensitivity analyses to ascertain the consistency of the model across different scenarios. Also, in order to increase the level of confidence in the results, complete statistical analysis and comparison with available methods would be recommended.

In order to offer a structure for similar works, one may extend the methodology employed in this study, encompassing data cleaning procedures, feature extraction elements, and the design of the stacking super classifier. This structure may then be altered depending on other tasks where cancer classification is the primary task or other similar problems involving genomic data analysis. Due to the models simplicity it can be used in conjunction with modern deployment tools like cloud services to allow classification on multiple devices like desktops, mobile phones and similar devices.

References

[1] Nair, J.R., et al. (2019). Novel Therapeutic Strategies for Soft Tissue Sarcomas. Advances in Experimental Medicine and Biology, 1257, 59-79.

[2] Integrated molecular characterization of soft tissue sarcomas using multi-omics data. Nature Communications, 12(1), 3611.

[3] Ren, T., et al. (2021). Prediction of clinical outcome and survival in soft-tissue sarcoma using radiomics and deep learning on preoperative MRI. European Radiology, 31(2), 749-760.

[4] LeBlanc, V.G., et al. (2017). Machine Learning on DNA-Methylation Data from High-Grade Sarcomas to Predict Relapse. Clinical Cancer Research, 23(20), 6288-6297.

[5] Liu, J., et al. (2018). An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell, 173(2), 400-416.e11.

[6] Movva, S., et al. (2019). Multi-platform profiling of over 2000 sarcomas: Identification of biomarkers and novel therapeutic targets. Oncotarget, 6(14), 12234-12247.

[7] van IJzendoorn DGP, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML, et al. (2019) Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLOS Computational Biology 15(2): e1006826.

[8] Alaoui EAA, Tekouabou SCK, Hartini S, et al. Improvement in Automated Diagnosis of Soft Tissues Tumors Using Machine Learning. Big Data Mining and Analytics, 2021, 4(1): 33-46.

[9] Schaefer IM, Fletcher CDM. Recent advances in the diagnosis of soft tissue tumors. Pathology. 2018 Jan;50(1):37-48. doi: 10.1016/j.pathol.2017.07.007. Epub 2017 Sep 23. PMID: 28950990.

[10] J. Juntu et al., 'Classification of Soft Tissue Tumors by Machine Learning Algorithms', Soft Tissue Tumors. InTech, Nov. 16, 2011. doi: 10.5772/27757.

[11] Wei, Y. , Gao, M. , Xiao, J. , Liu, C. , Tian, Y. and He, Y. (2023) Research and Implementation of Cancer Gene Data Classification Based on Deep Learning. Journal of Software Engineering and Applications, 16, 155-169. doi: 10.4236/jsea.2023.166009.

[12] Maria Ali, Muhammad Nasim Haider, Saima Anwar Lashari, Wareesa Sharif, Abdullah Khan, Dzati Athiar Ramli, Stacking Classifier with Random Forest functioning as a Meta Classifier for Diabetes Diseases Classification, Procedia Computer Science, Volume 207, 2022, Pages 3459-3468, ISSN 1877-0509.

[13] Alexandropoulos, SA.N., Aridas, C.K., Kotsiantis, S.B., Vrahatis, M.N. (2019). Stacking Strong Ensembles of Classifiers. In: MacIntyre, J., Maglogiannis, I., Iliadis, L., Pimenidis, E. (eds) Artificial Intelligence Applications and Innovations. AIAI 2019. IFIP Advances in Information and Communication Technology, vol 559. Springer, Cham.

[14] Dzeroski, S.,ˇ Zenko, B. Is Combining Classifiers with Stacking Better than Selecting the Best One?. Machine Learning 54, 255–273 (2004).ˇ [11] Cao,S.,et al. (2021).

[15] Taylor, B. S., Schultz, N., Hieronymus, H., Gopalan, A., Xiao, Y., Carver, B. S., Arora, V. K., Kaushik, P., Cerami, E., Reva, B., Antipin, Y., Mitsiades, N., Landers, T., Dolgalev, I., Major, J. E., Wilson, M., Socci, N. D., Lash, A. E., Heguy, A., ... Gerald, W. L. (2010). Integrative Genomic Profiling of Human Prostate Cancer. Cancer Cell, 18(1), 11-22.

[16] The Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. Nature 541, 169–175 (2017).

[17] cBioPortal for Cancer Genomics. (2024, July, 15). Adult Soft Tissue Sarcomas (TCGA, Cell 2017). [Online]. Available: https://www.cbioportal.org/study/summary?id=sarc_tcga_pub

[18] cBioPortal for Cancer Genomics. (2024, Sep, 22). Soft Tissue and Bone Sarcoma (MSK, Nat Commun 2022). [Online]. Available: https://www.cbioportal.org/study/summary?id=sarcoma_msk_2022