AVE Trends in Intelligent Computer Letters

Improving Cancer Subtype Classification in High-Dimensional Gene Expression Data via PCA and Machine Learning

Rejwan Bin Sulaiman[1,*], **Noman Javed**[2]

[1]Department of Computing Science and Technology, Northumbria University, Newcastle upon Tyne, England, United Kingdom.
[2]School of Computing, Engineering and Physical Science, University of the West of Scotland, Paisley, Scotland, United Kingdom.
rejwan.sulaiman@northumbria.ac.uk[1], b01794811@studentmail.uws.ac.uk[2]

*Corresponding author

Abstract: The problem of predicting cancer subtypes is crucial to contemporary research on cancer because data on gene expression are used to reveal the patterns that distinguish among different subtypes. The paper will discuss machine learning models, namely, Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF), to be used to classify the subtypes of LGG and BRCA cancer. Among the key issues handled are high-dimensional data and small sample sizes. To prevent overfitting and enhance model generalisation, Principal Component Analysis (PCA) was used as a dimensionality-reduction method. The models were assessed using metrics such as accuracy, macro-F1 score, ROC-AUC, and confusion matrices, and, to achieve reliable results, nested cross-validation was employed. These results demonstrate that PCA enhances model generalisation, with the SVM+PCA combination being the most accurate and robust across both datasets. Stratified sampling and class weighting were implemented to address class imbalance, especially in the BRCA dataset. Integration of PCA not only minimised overfitting and enhanced model stability but also simplified the interpretation of results, without using complex, multifaceted omics data or deep learning techniques. Although the findings were encouraging, the research could be improved in the future by considering more sophisticated resampling methods and omics approaches to further increase predictive value.

Keywords: Logistic Regression (LR); Support Vector Machines (SVM); Biological Data; Random Forests (RF); Cancer Subtypes Prediction; Machine Learning Models; Gene-Expression Data.

Cite as: R. B. Sulaiman and N. Javed, "Improving Cancer Subtype Classification in High-Dimensional Gene Expression Data via PCA and Machine Learning," *AVE Trends in Intelligent Computer Letters*, vol. 1, no. 4, pp. 208–231, 2025.

Journal Homepage: https://www.avepubs.com/user/journals/details/ATICL

Received on: 03/11/2024, **Revised on:** 19/12/2024, **Accepted on:** 20/03/2025, **Published on:** 07/12/2025

DOI: https://doi.org/10.64091/ATICL.2025.000235

1. Introduction

1.1. Background

Machine learning enables effective cancer subtype prediction by analysing high-dimensional biological datasets and uncovering complex patterns that may not be detectable through traditional statistical techniques. In oncology, gene-expression profiles often contain thousands of genes measured across a relatively small number of patients, creating a "high-p, low-n" scenario

[2]. Biological data is rich in useful information, but can be noisy and difficult to analyse on its own without computational tools. Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF) are machine learning models that provide useful, interpretable methods for identifying subtype-specific patterns in complex data [3]. The models play a significant role in enabling early diagnosis and an individualised treatment plan by learning from large feature spaces [4]. Cancer is a highly heterogeneous disease; in other words, even patients with the same clinical diagnosis can vary in their molecular profiles [6]. These types of molecular variability are likely to define various subtypes, each with an assigned prognosis and treatment options. It cannot afford to rely on a single biomarker or clinical variable to classify the subtype, according to Cai et al. [49]. A more rigorous approach can provide a better understanding of tumour biology, revealing elaborate patterns of gene expression. Still, machine-learning models can recognise subtle differences among disease subgroups. The dimensionality of gene-expression data is high, which poses several issues, including overfitting, increased computational time, and the inability to extrapolate to new data [10]. Dimensionality-reduction algorithms, such as Principal Component Analysis (PCA), can be used to address these problems by reducing many correlated variables into fewer meaningful components [11]. The process can minimise noise, focus on significant differences in the data, and improve classifier performance.

According to Acharya and Mukhopadhyay [8], the main factors that can help ensure the reliability, reproducibility, and clinical relevance of machine-learning models for cancer prediction are adequate preprocessing, dimensionality reduction, and comprehensive validation [12]. Cancer subtype prediction is a crucial aspect of modern oncology, as different subtypes can differ greatly in terms of prognosis, treatment options, and response to therapy [14]. Among the various data types used in cancer research, gene-expression profiling is one of the most commonly generated and accessible sources of molecular information [15]. In these datasets of gene-expression information, the number of features describing gene activity in tumours is often in the thousands, providing useful information but also posing considerable computational challenges for modelling [16]. In this context, machine-learning methods provide effective tools for recognising patterns within high-dimensional expression data and supporting more accurate subtype classification [49]. The LGG and BRCA datasets used in this paper, however, are provided in a simplified, preprocessed format containing only gene expression data [20]. To perform proper omics integration, separate raw omics layers would need to be downloaded, harmonised, matched by sample IDs, normalised independently, and processed into block matrices [21]. These steps demand significant computational power and storage resources [24]. These requirements extend beyond the scope of a single-student MSc in Data Science. As omics datasets are integrated into models, these models are computationally intensive and often require GPU acceleration or high-performance computing environments [26]. Frameworks and multiblock integration frequently involve millions of parameters, making them unsuitable for execution on standard laptops [27].

The aim of this paper is not to replicate large-scale, multi-layer biological modelling but to evaluate the performance of classical machine-learning models in a more manageable setting that remains scientifically meaningful [28]. Lastly, multi-omics research demands strict handling of missing modalities, batch effects, and cross-platform variability [34]. In real-world TCGA cohorts, many patients lack complete data across omics assays. Handling such incomplete blocks requires advanced imputation strategies or specialised models designed for partial-view learning—again, far beyond the requirements and scope of an MSc-level paper [36]. ML in gene-expression data, which is both appropriate for the available resources and representative of how many machine-learning studies in oncology begin [37]. Classical models such as Logistic Regression, Support Vector Machines, and Random Forests provide strong and interpretable baselines for gene-expression classification [38]. However, due to the extremely high dimensionality of transcriptomic data, these classifiers can overfit when used directly [1]. To address this, Principal Component Analysis (PCA) is employed as a dimensionality-reduction step. PCA transforms thousands of correlated genes into a compact set of uncorrelated components, capturing the most informative variance, improving model generalisation, and reducing computational complexity [8]. By focusing on PCA-enhanced classical machine learning, this study achieves a balance between feasibility, interpretability, and academic relevance, while avoiding the heavy computational burden and methodological complexity associated with omics data integration [41]. The selected approach is therefore appropriate for the paper's resources and aligns with widely accepted practices in transcriptomic machine-learning research [44].

1.2. Research Questions and Objectives

- **RQ1 (Effectiveness):** Do baseline models (Logistic regression (LR), support vector machine (SVM), and random forest (RF) predict the omics, such as LGG and BRCA subtypes?
- **RQ2 (Interpretability):** Do integrative pipelines of models (LR+PCA, SVM+PCA, and RF+PCA) produce stable and biologically coherent feature attributions and per-omic contribution patterns compared to interpretable baselines?
- **RQ3 (Robustness):** How robust are integrative models to real-world "mosaic" scenarios for LGG and BRCA subtype detections?

Although cancer is influenced by multiple molecular processes, many practical machine-learning studies begin with omics datasets, particularly gene-expression profiles, because they are widely available, well-characterised, and computationally manageable. While omics datasets integration has the potential to capture interactions across genomics, methylation, and transcriptomics, such approaches require multiple complete data layers, complex preprocessing, and substantial computational resources—conditions that are not met in the present study [47]. Instead, this paper focuses on addressing a key challenge inherent to gene-expression data: extremely high dimensionality relative to the number of samples. This imbalance can cause classical machine-learning models to overfit and limit their ability to generalise.

1.3. Research Question

How effectively can machine-learning models classify LGG and BRCA cancer subtypes, and does PCA improve their predictive performance? To develop and evaluate machine-learning pipelines that integrate heterogeneous omics to improve cancer subtype classification while maintaining interpretability and robustness relative to single-omics baselines:

- Implement baseline single-omics classifiers (e.g., Random Forest, SVM, Logistic Regression) for LGG and BRCA omics prediction.
- Build integrative pipelines (e.g., dimensionality reduction via PCA; latent-factor and multiblock projection methods) and compare against baselines.
- Evaluate accuracy**,** macro-F1, ROC-AUC, and confusion matrices (including missing-modality scenarios) under nested cross-validation.

To mitigate this issue, dimensionality-reduction approaches such as Principal Component Analysis (PCA) are increasingly used to extract the most informative structure from gene-expression data before classification. PCA reduces noise, reveals dominant patterns, and produces a more compact feature representation, enabling models such as Logistic Regression, Support Vector Machines, and Random Forests to perform more effectively. Considering these issues, the research aims to evaluate the impact of adding PCA to standard machine-learning pipelines on improving predictive accuracy and stability for LGG and BRCA subtypes. Rather than pursuing multi-omics integration, the research will aim to reinforce classical single-omics modelling through extensive preprocessing, dimensionality reduction, and systematic evaluation.

1.4. Problem Definition

There are many cancer gene-expression datasets with thousands of features and only a few samples, making them prone to overfitting when trained with classical machine-learning models. The high-dimensional, low-sample structure renders models such as Logistic Regression, SVMs, and Random Forests challenging to generalise appropriately unless proper dimensionality reduction is performed. A large amount of data can add redundant or noisy information, i.e., using the complete set of features may decrease classifier performance and make computation more expensive. Challenges in gene-expression modelling are class imbalance, particularly in BRCA subtypes, where some classes are substantially under-represented. Imbalanced data can bias models toward majority classes and reduce macro-F1 performance, which is crucial for fair evaluation across all subtypes. Interpretability also remains important in biomedical applications. While classical machine-learning models are more transparent than deep learning, high dimensionality still makes it difficult to understand which features or components drive model behaviour. PCA provides a way to summarise variance across thousands of genes into a smaller number of components, making the data easier to analyse and reducing noise before classification. These challenges highlight the focus of this paper: evaluating whether PCA-based dimensionality reduction can enhance the accuracy, robustness, and interpretability of classical machine-learning models for LGG and BRCA subtype prediction using high-dimensional gene-expression data.

1.5. Aim and Scope

The gaps are addressed by systematically comparing machine-learning models with PCA-enhanced versions using LGG and BRCA gene-expression datasets. The study employs critical evaluation practices, including multiple train–test partitions and k-fold cross-validation, to assess stability and generalisation. By examining how PCA transforms high-dimensional data and influences model accuracy, macro-F1, and ROC-AUC, the paper provides practical insight into the role of dimensionality reduction in cancer subtype prediction. Due to limited sample size and high dimensionality of gene-expression data, researchers focused on single-omics transcriptomic profiles and classical machine-learning techniques rather than deep-learning or multi-omics integration, because these approaches would increase the risk of overfitting, require more computational resources, and complicate biological interpretation beyond the scope and resources of this study. Deep learning methods may not outperform other machine learning methods in analysing genomic studies – shows that, even with large genomic datasets, deep learning does not always outperform classical methods in genomic classification tasks due to data structure and sample size constraints [46]. A survey of single-omics data mining methods in cancer research discusses challenges in integrating omics datasets, including the complexity of combining heterogeneous omics sources and the statistical pitfalls that arise when sample sizes are limited [50].

1.5.1. Scope

This paper uses publicly available, de-identified gene-expression datasets from The Cancer Genome Atlas for LGG and BRCA. The study relies entirely on secondary analysis of open-source data; no new data collection, patient recruitment, or identifiable clinical information is involved. Only the transcriptomic (RNA-seq) expression matrices provided in processed form are used. The study evaluates classical machine learning models for cancer-subtype prediction using gene expression data. Two categories of models are compared:

- **Baseline Models:** Logistic Regression, Support Vector Machine, and Random Forest applied directly to high-dimensional features.
- **PCA-Enhanced Models:** LR+PCA, SVM+PCA, and RF+PCA, using PCA to reduce dimensionality before classification.

Performance is assessed using accuracy, macro-F1, ROC-AUC, and confusion matrices. Multiple train–test splits and k-fold cross-validation are used to examine stability and generalisation. The deep learning, multi-omics integration, or advanced latent-factor models are not included, as the study focuses specifically on machine-learning workflows for datasets. The main objective is supervised classification of LGG and BRCA cancer subtypes based on gene expression profiles, and it does not explore other clinical outcomes, such as patient survival prediction, treatment response, or biomarker discovery. Moreover, the focus is on evaluating how PCA-based dimensionality reduction affects the performance of machine-learning models in subtype classification.

1.5.2. Delimitations

The study is confined to single-omics gene-expression data, as matched multi-omics layers (e.g., methylation, CNA, miRNA) are unavailable in a harmonised format within the paper's timeframe. External validation using fully independent datasets is also not feasible due to time and data constraints. The research has used consistent cross-validation protocols, cross-validation across multiple scales, leakage-free preprocessing (scaling and PCA are applied only to the training folds), and a clear description of all parameters and evaluation metrics to ensure reliability within the constraints. The models are contrasted with standard hardware that limits the usage of computationally intensive methods.

1.5.3. Research Gap and Rationale

The high-dimensional data from gene expression pose significant challenges for predicting cancer outcomes; however, several gaps remain in the application and testing of classical machine-learning methods in this field. Numerous papers report high accuracy but tend to explicitly operate on thousands of gene features, neglecting the underlying challenge of small-sample, high-dimensional data. The strategy may result in overfitting, inept cross-validation, and overly optimistic execution estimates, especially when no dimensionality reduction or appropriate cross-validation is applied. A systematic review of dimensionality-reduction approaches, e.g., Principal Component Analysis (PCA), within the scope of ML pipelines is needed, especially for predicting LGG and BRCA subtypes. The key ML models, such as Logistic Regression, Support Vector Machines, and Random Forests, have been extensively applied. Yet, their performance is compared under stable overall evaluation conditions with and without dimensionality reduction. The knowledge gap in this case is a lack of understanding of the situations in which PCA is most applicable, the amount of variance it should retain, and the variability in model performance across different data partitions. In particular, stability across multiple train–test splits and cross-validation folds is rarely emphasised, even though robustness is essential for reliable biomedical prediction. Third, interpretability in high-dimensional settings remains a challenge. While classical ML models are inherently more transparent than deep learning, the presence of tens of thousands of correlated gene features makes direct feature-level interpretation impractical. PCA provides an opportunity to summarise key patterns in the data, but empirical studies that examine how PCA affects model interpretability, variance structure, and downstream classification performance are limited.

1.6. Significance and Expected Contributions

This study makes a meaningful contribution to the field of cancer subtype prediction by systematically evaluating how dimensionality reduction through Principal Component Analysis (PCA) influences both the performance and generalisation of classical machine-learning models applied to high-dimensional gene-expression data. While many existing studies apply machine-learning algorithms directly to transcriptomic datasets, far fewer conduct a controlled and reproducible comparison between baseline models and PCA-enhanced pipelines. By isolating the impact of PCA on Logistic Regression, Support Vector Machines, and Random Forest. Forests, a clear empirical basis for understanding when dimensionality reduction leads to gains in (i) accuracy, macro-F1, and ROC-AUC, and when classical models perform comparably without feature compression. (ii) The work also surfaces methodological guidance on the practical use of PCA for stabilising model behaviour, reducing dimensional noise, and ensuring efficient computation in settings where datasets are high-dimensional but sample sizes remain limited. This guidance is particularly relevant for researchers working with datasets or similarly structured biomedical data who require models that balance performance, interpretability, and computational feasibility. The paper directly addresses the high-dimensional, low-sample challenge common in cancer datasets. The experiment used different train-test splits and k-fold cross-validation to evaluate how PCA reduces overfitting, improves stability, and makes the training process more efficient. The findings are applied to develop powerful, robust machine-learning processes in biomedical studies, where it is critical to ensure generalisation and to define transparent procedures. The practice aligns with the objectives of the MSc Management with Data Analytics program, as it helps acquire essential skills in pre-processing data, classical model construction, performance assessment, and research ethics and reproducibility. The application to real-world TCGA data provides a useful experience in addressing the challenges of biomedical analytics, including class imbalance, high-dimensional data, and the importance of a well-designed experimental design.

1.6.1. Artefacts and Reproducibility

Reproducibility is a core principle throughout the experimental pipeline. All preprocessing steps, including scaling and PCA, are applied solely within the training folds to prevent data leakage. Random forest and controlled data partitions ensure consistent reruns, and all metrics are transparently reported for each variable configuration. While the study focuses on machine-learning models rather than deep learning, the reproducibility standards followed align with best practices that are becoming the norm in biomedical machine-learning research [51].

1.6.2. Ethical Considerations

The paper uses only public, anonymised datasets; there is no human subject interaction. University policies on data governance, research integrity, and academic honesty are followed. Researchers avoid handling identifiable data, respect dataset licences, and document decisions to support transparent, reproducible ML practice.

2. Literature Review

2.1. Overview of Cancer Subtype Prediction

Over the past decade, oncology has increasingly adopted a data-driven approach to discovery and decision-making. The rapid growth of data has shifted research from traditional hypothesis-driven experiments to data-intensive analytical science. The growing complexity and heterogeneity of data have made it clear that “one-size-fits-all” treatment strategies are no longer sufficient. Precision oncology now uses large-scale data analytics to develop personalised treatment plans [43]. As Baptiste et al. [29] present, this transformation has been propelled by integrating genomic, epigenetic, and environmental factors into computational models that predict patient-specific outcomes. Massive data: the dissemination of big data. Projects like the Human Genome Project and the 1000 Genomes Project foreshadowed data-driven medicine (by creating large, high-dimensional datasets). Such efforts led to the development of medical datasets that are now important for research. Machine learning (ML) is currently an essential tool in modern biomedical research, as it can model intricate, non-dimensional relationships that cannot be achieved with traditional statistical methods. The problem of predicting cancer subtype with high p-values and low n means there are hundreds of thousands of features but only a few samples, a typical feature of genome-wide gene expression data. This type of data structure is challenging for traditional modelling, and it is worth noting that ML plays a critical role in defining the vital patterns that distinguish clinically significant tumour subtypes.

2.2. Machine Learning in Biomedical Research

The main problem in data science, as it is perceived nowadays, is not necessarily high classification accuracy, but rather the construction of pipelines that are not only computationally efficient but also reproducible. In this case, algorithms require approaches such as regularisation and variance reduction, and to prevent overfitting, appropriate judgment is required to ensure the number of features is greater than the sample size. It is this problem that spawned machine learning in genomics; dimensionality reduction, well-regularised models, and strong cross-validation are of key interest. Logistic Regression, Support Vector Machines, and Random Forests, which are traditional machine learning techniques, have been reported as reliable for analysing gene expression. Such methods are not as sophisticated as deep learning models, yet they can compete with them, trained with suitable preprocessing. This can be illustrated by applying Logistic Regression (LR) with L2 regularisation, which is appropriate for sparse, high-dimensional data. Support Vector Machines (SVMs), particularly with linear kernels, are widely used in cancer subtype prediction because they maximise the margin between classes, thereby improving generalisation to new data. Random Forests complement these methods by providing model robustness, built-in feature ranking, and resilience to noise. Machine learning models require strategies to counteract the “Curse of Dimensionality,” in which irrelevant or redundant features obscure the true structure of the data. This is where dimensionality reduction techniques, especially Principal

Component Analysis (PCA), play a transformative role. A useful way to organise the literature is by fusion strategy, because where and how fusion happens strongly influences both performance and engineering complexity. Early fusion concatenates features across assays after per-omic preprocessing and then trains a conventional classifier. It is easy to implement and can be competitive when the signal is strong and well-aligned across modalities. Late fusion trains a separate model per assay and combines outputs via voting, stacking, or calibrated averaging, which can be robust when modalities disagree or are missing at inference time. Middle fusion learns a joint representation and an embedding before classification, using tools such as factor models, supervised multiblock projections, or deep/graph architectures. The appeal of middle fusion is that the model can discover and exploit cross-omic correlations while being guided by the task loss, thereby tuning the representation to the downstream objective. Most reports that convincingly beat strong baselines fall into this middle-integration category, particularly when evaluation controls for leakage and ensures that preprocessing is fit only within training folds [49].

2.3. Dimensionality Reduction in Genomic Data

Dimensionality reduction is a cornerstone of ML pipelines for bioinformatics. Gene-expression datasets typically contain thousands of correlated genes, many of which contribute minimal signal for subtype prediction. PCA reduces high dimensionality by capturing the dataset's variance using a smaller number of orthogonal components. They are linear mixtures of genes that characterise important patterns of variance, and downstream models can be run on a lower-dimensional, lower-noise reduced space [23]. PCA has three crucial uses in cancer subtype prediction: reducing noise, accounting for biological variability, measurement noise, and batch effects in gene-expression data. The technique is useful because it eliminates low-variance variables, which may be noisy rather than adding biological information. With PCA-transformed features, classifiers can use fewer parameters and are therefore less prone to overfitting, particularly in high- and low-sample-size environments and in high-feature environments. It also reduces computation costs, memory usage, and runtime. This is significant in resource-limited environments, as encountered in an MSc paper. PCA is applied to improve model performance. One such example is that, when cancer data is used with PCA followed by SVM classification, generalisation is better than when the raw data is used. PCA on Random Forests can also be used to address situations in which a large number of overlapping gene groups can make decision-making erratic [30]. Still, PCA is associated with interpretational challenges, as the most significant components are mathematical constructs that do not directly correspond to biological attributes. To be specific about the effects of the transformed features on the classification outcome, one should report the PCA loadings and the variance of each component. The most popular machine learning models, including Logistic Regression, Support Vector Machines, and Random Forests, are widely used in cancer subtype prediction because they can handle high-dimensional gene expression data, a complex dataset. Logistic Regression continues to contribute to biomedical prognostication by providing decision-making capabilities and accounting for probabilistic outcomes.

LR is particularly effective in the L2 regularisation scenario, as it prevents overfitting and enables generalisation [30]. Large coefficients are penalised by the regularisation, encouraging the model to place greater weight on important features. Previous studies have shown that LR performs competitively when the transformed feature space is approximately linearly separable, making it a strong baseline for models that require transparency and statistical stability in complex biomedical settings. Support Vector Machines have a long history of success in gene-expression modelling. Linear SVMs are especially well-suited for cancer subtype prediction because they optimise a margin-based objective that tends to generalise well even when the number of features vastly exceeds the number of samples. Their reliance on support vectors rather than the full dataset makes them resistant to noise and robust under sparse, high-dimensional conditions. SVMs avoid many of the instability issues observed in less constrained models, and empirical evidence consistently places them among the best-performing algorithms for microarray and RNA-seq classification tasks. Random Forests address limitations associated with linear decision boundaries by capturing nonlinear interactions among features. By combining decorrelated decision trees, RF models offer resilience to noise and natural robustness against overfitting. Random Forests (RFs) also provide feature-importance rankings, which makes them valuable for exploratory biomedical studies where identifying influential variables is key. While RFs can perform well without dimensionality reduction, extremely high-dimensional data, such as transcriptomic matrices containing tens of thousands of genes, can lead to instability in tree splits. When a technique such as PCA is used, it can produce more consistent partitions and minimise variance between model runs, thereby increasing overall model reliability [25].

2.4. Class Imbalance in Cancer Data

Though machine learning methods are advantageous, several challenges remain that gene-expression modelling struggles with and that influence the decisions made during the research. They tend to be significantly smaller than the datasets used in gene expression analysis and can readily lead to overfitting that cannot be remedied by dimensionality reduction or regularisation. The effect of such overfitting can be low model generalisation. The given issue is often addressed in genomic research, where Principal Component Analysis (PCA) is useful for feature selection and regularisation [33]. There are underrepresented subtypes in the prediction of cancer subtypes, for example, in a database such as the BRCA database. Such an imbalance can cause bias in the models, in which the models blow out most of the classes and ignore the minority classes. In this manner, measures such as macro-F1 and ROC-AUC are preferred for assessing model performance, as they assess the predictive ability across all classes, not just the dominant ones. Information loss due to data leakage is another common type of information loss that occurs when preprocessing procedures are applied to all the input data before partitioning it into training and test sets, such as scaling or feature selection.

It can lead to overfitting, which is why it is important to preprocess only the training set. Use of the guidance to prevent leakage will make the model evaluation valid. PCA is employed to reduce the dimensionality of gene-expression data and identify and systematise the most valuable elements. PCA identifies and removes the main sources of variation and noise, helping stabilise models and eliminate overfitting, particularly with massive gene-expression data [5]. The samples are balanced so that the subtype ratios are equal between the training and test conditions. It mitigates bias and ensures sound performance estimates. Preprocessing of the entire training set is performed to prevent data leakage and ensure reproducibility, as recommended in the field. Some measures employed to assess the models include accuracy, macro-F1, ROC-AUC, and confusion matrices. These measures provide an in-depth evaluation of the models' performance in real-world contexts, especially with uneven data sets. PCA helps enhance the stability and performance of a model by producing more manageable data networks from high-dimensional data. PCA can be used to reduce dimensionality, but it is not optimised for class separation. The elements obtained in PCA summarise the variance rather than focusing on class distinction. Thus, the optimal number of components is determined using cross-validation to ensure the model is trained on the best features.

2.5. Existing Work in Cancer Subtype Prediction

A key aspect of machine learning in genomics is the importance of proper evaluation methodology. Early studies on cancer subtype prediction often overestimated their performance due to issues such as improper cross-validation or feature selection performed before data splitting [7]. Cancer subtype prediction plays a vital role in personalised medicine, as it helps identify tailored therapeutic strategies based on a patient's cancer's molecular features. Over time, various machine learning techniques have been applied to classify cancer subtypes using genomic data, such as gene expression profiles, methylation data, and somatic mutations.

2.5.1. Traditional Approaches in Cancer Subtype Prediction

In previous research on cancer classification (subtype), major methods included hierarchical clustering, principal component analysis (PCA), and unsupervised learning. This was done to the gene expression data using these methods to classify the samples into subtypes according to gene expression similarities. One such example is the study by Greenacre et al. [31], which, after performing hierarchical clustering of expression data for breast cancer genes, identified subtypes subsequently shown to have varying prognostic and therapeutic implications. Although these techniques were useful for clustering cancer samples, they struggled with high-dimensional data (an abundance of features and few samples) and may have overfit the data. In response to these problems, researchers began using stronger machine learning models, particularly supervised learning algorithms, which better supported high-dimensional data and provided greater generalisation [22].

2.5.2. Machine Learning Models for Cancer Subtype Prediction

Cancer subtype prediction has been performed using more sophisticated machine learning models, including Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression (LR). Specifically, SVMs are widely used for cancer classification with high-dimensional data and even with small sample sizes. As a case example, Lu et al. [45] studied breast cancer subtypes using SVM on gene expression data, and the results were quite good. Ensemble methods like random forests have also gained popularity for their robustness and ability to handle a large number of features. Research papers such as those by Chaudhary et al. [27] have shown that Random Forests can achieve high classification performance on cancer genomic data, often surpassing that of conventional methods. Nevertheless, these models are effective for classification but may be difficult to interpret, which is essential for understanding the biological mechanisms underlying cancer subtype differentiation. To curb this, scholars have resorted to dimensionality reduction tools such as Principal Component Analysis (PCA), which, while improving model interpretability, preserves model performance.

2.5.3. Dimensionality Reduction and Feature Selection

Dimension reduction algorithms such as PCA are a common method of data analysis in machine learning, as they make it easier to work with gene expression data and help avoid overfitting. PCA is a method that attempts to convert the original variables into fewer variables that capture the greatest variation in the data. Such components can be given to machine learning models. A combination of PCA and machine learning models would address the generalisation and interpretability issues identified by Meshoul et al. [39]. PCA will also reduce feature complexity when predicting cancer subtypes, thereby reducing the risk of overfitting while retaining the most important biological drivers in the gene expression data. The usual practice in cancer classification has shifted to using PCA, SVMs, or random forests. As shown in a study by Flores et al. [24], the SVM + PCA combination is highly effective for categorising lung cancer into subtypes. PCA was used to eliminate redundant features and focus on the most informative ones.

2.5.4. Class Imbalance in Cancer Data

Class imbalance, an issue in cancer subtype prediction, is where some cancer subtypes are underrepresented in the data, leading the model to predict the majority class. Several approaches, such as resampling methods (e.g., SMOTE - Synthetic Minority Over-sampling Technique) and class-weight modulations in models such as SVM and Random Forest, have been employed to address this. Sugianto and Wahyuningsih [9] proposed SMOTE, which has become the most popular method for class balancing, creating synthetic examples of minority classes to balance the majority classes. The techniques enhance the model's performance by providing a more balanced representation of the minority classes. Sakri and Basheer [40] used class balancing and feature selection in gene expression data to improve the predictive performance for breast cancer subtypes, achieving higher-quality metrics, such as F1-score and ROC-AUC, than in imbalanced data. Similarly, the issue of class imbalance needs to be addressed when predicting the subtypes of BRCA and LGG cancers, as many of these cancers have far fewer samples.

2.5.5. Interpretability of Models in Cancer Subtype Prediction

Machine learning algorithms such as SVMs and Random Forests can also be useful for predicting cancer subtypes, which are often considered black-box models [39]. This presents some difficulties in clinical use, where interpretability would be vital for understanding how to increase or decrease cancer progression and respond to treatment. Explainable machine learning methods, such as SHAP (SHAPley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations), have been applied to enhance interpretability. Such methods provide insights into the premises of expert models by assigning a score to each characteristic, indicating its contribution to the model outcome [35]. Nicora et al. [19] demonstrated the use of SHAP values to interpret the predictions of Random Forest models in cancer classification, enabling researchers to identify which genes have the greatest influence on predicting cancer subtypes. PCA itself can provide interpretability by summarising the most important variance in the data, offering insights into which features (genes) contribute most to cancer subtype differentiation [32].

3. Methodology

3.1. Research Design

This paper evaluates machine-learning strategies for omics data integration to predict LGG and BRCA cancer subtypes as a supervised classification task. The design is deliberately simple and reproducible: I standardise preprocessing across datasets, compare a small set of well-established integration families under the same splits, and use leakage-free model selection with cross-validation. The goal is to produce a pipeline that is easy to implement but still aligns with best practices highlighted in recent surveys. I restrict the comparison to three integration strategies that cover the main middle-fusion paradigms while keeping engineering effort low:

- Early-fusion regularised linear baseline (LR, SVM, and RF).
- PCA + simple classifier using (learn joint latent factors, then Logistic Regression).

3.2. Data Collection

I work with publicly available, anonymised omics data cohorts with labelled subtypes of LGG and BRCA, focusing on tasks where each sample has at least one omics layer and a valid class label. This keeps the study within a single analytics domain (supervised classification) and avoids biological interpretation:

- Include samples with a known subtype label for both LGG and BRCA cancer subtypes.
- Exclude male samples from the BRCA dataset and zero columns.
- Without labels, have only one class value; also specify the other class as 'other cancer' for the BRCA dataset.

Datasets are public and de-identified; there is no participant contact or intervention. All files and derived Tables are versioned, and a simple data dictionary documents features per omic and any filtering steps.

3.3. Preprocessing and Feature Engineering

To prevent leakage, every preprocessing step is fit within the training portion of each cross-validation split and applied to the corresponding validation/test split, using saved parameters (Figure 1).

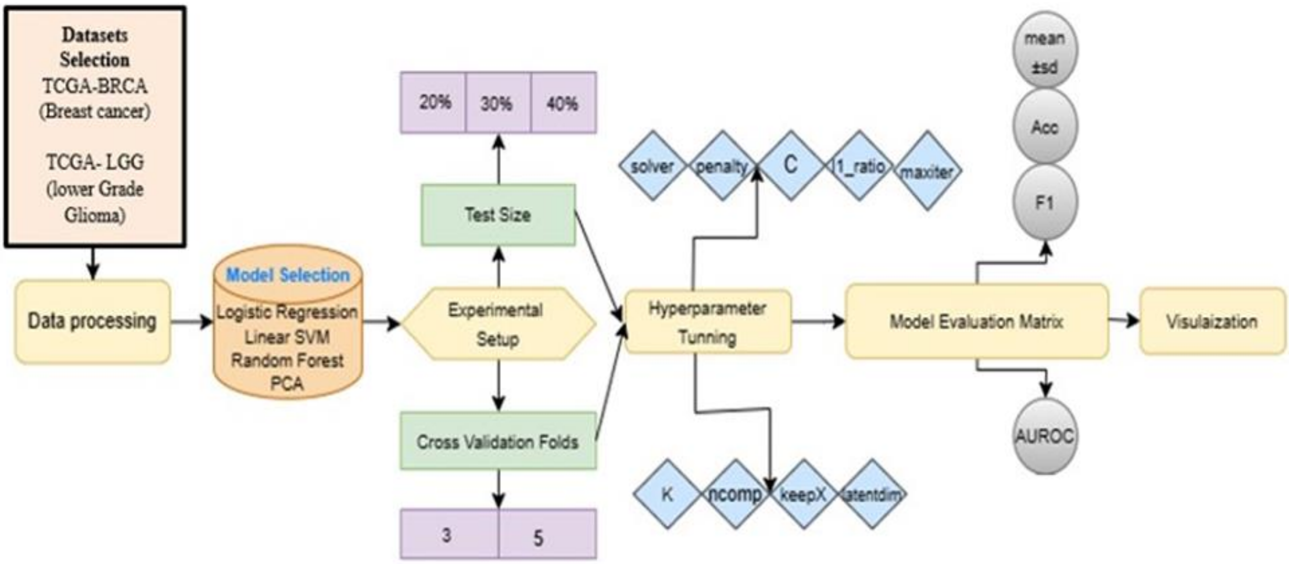


Figure 1: Methodology for integrating machine learning for accurate cancer type prediction

3.3.1. Per-Omic Cleaning and Scaling

- Remove features with zero variance.
- Winsorize extreme outliers (e.g., 0.5%/99.5%).
- Standardise each feature to zero mean/unit variance on the training fold; apply the same scaler to val/test.

Where batch variables are available, I prefer model-level robustness (representation learning) over heavy domain-specific corrections. As a sensitivity check, I may compare it to a simple empirical-Bayes adjustment, while keeping the main pipeline domain-agnostic.

3.3.2. Feature Filtering and Dimensionality Reduction

- Keep top-kkk features per omic by robust variance (or mutual information with the class) computed on training only.
- For early-fusion baselines, optionally add PCA per omic (fit on training only) if feature counts remain very high.

3.4. Interpretability of Models

Interpretability in machine learning, especially in biomedical applications, is crucial for ensuring that the findings can be trusted and understood by experts in the field. In this study, while classical machine-learning models such as Logistic Regression (LR), Support Vector Machines (SVM), and Random Forests (RF) offer greater transparency than deep learning models, the high dimensionality of gene-expression data poses challenges for feature interpretation. To address the challenges of high-dimensional gene-expression data, PCA was used as a dimensionality-reduction technique to transform the dataset into a smaller set of uncorrelated components, simplifying the analysis. PCA poses interpretability-related issues of its own: it produces components that are linear combinations of the original features and are not directly related to any particular biological characteristics. Although PCA can reduce the dimensionality and noise, it complicates the task of explaining the biological meaning of each principal component. To enhance interpretability:

- **Component Loadings:** The loadings of the main constituents were used to identify the genes most related to each component. The variables help identify the most important features (genes) in each of the main aspects and, consequently, the classification as a whole.

- **Percentage of Explaining Variance:** The percentage of variance explained by each principal component is reported, providing information on how much of a biological signal (rather than noise) is explained by each component [31].
- **Biological Significance:** Another step is to cross-reference the top contributing genes of each major component with known biological information or databases (e.g., Gene Ontology, KEGG pathways) to determine their biological relevance.
- **Visualisations:** Biplots and loading plots were computed to enhance interpretability and demonstrate the contribution of each principal component to model classification. These diagrams can be used to put into perspective the features and elements that would influence the model's decisions.

3.4.1. Overfitting

Overfitting, especially with high-dimensional data such as gene expression data, is a frequent problem when the feature-to-sample ratio is high. In this work, overfitting was observed in the baseline models (Logistic Regression, SVM, and Random Forest), which performed well on the training data but significantly worse on the test data. This shows that the amount of noise learned by the models was related to the noise and not the pattern. To fight overfitting, the following means were employed:

- **Dimensionality Reduction:** Principal Component Analysis (PCA) has been used to reduce the dataset's feature space, retain the most significant variations, and eliminate noise and redundant features. This makes the model more general by focusing on the most relevant information.
- **Cross-Validation:** Cross-validation with splits of three or more was performed to assess the stability and performance of the models across different data splits. This approach helps prevent overfitting, in which the model generalises poorly across different subsets of the data and is overly affected by the training and test sets.
- **Regularisation:** In models such as Logistic Regression, L2 regularisation penalises large weights. This will also prevent the model from accepting very minor variations in the data and will focus only on the most significant features.
- **Class Weights Adjustment:** Class imbalance, especially across BRCA subtypes, was addressed by adjusting class weights. This ensured that underrepresented subtypes received proper attention in the model, thereby increasing its ability to categorise minority classes.

3.5. Model Training and Reproducibility

3.5.1. Languages/Libraries

- Python, Scikit-learn (preprocessing, Logistic/SVM/RF, metrics), pandas, numpy.
- Deterministic seeds for numpy/sklearn.
- All artefacts (scalers, PCA) are saved per fold for exact reproducibility.

These practices are repeatedly recommended in ML-centric omics overviews. I will use nested cross-validation to avoid optimistic bias:

- Accuracy, macro-F1, ROC-AUC, and confusion matrices with 20%, 30%, and 40% testing sizes.
- All preprocessing (scaling, PCA, feature filters) is fit on the inner/outer training data only and applied to the validation/test partitions using the stored parameters.
- Repeat the outer CV with 3- and 5-fold splits across multiple random seeds to estimate variability.

In this study, nested cross-validation (CV) was applied to obtain unbiased estimates of model generalisation performance while rigorously tuning hyperparameters, ensuring that hyperparameter selection and performance evaluation remain separate and avoiding over-optimistic bias arising from overlapping tuning and testing data sets [42]. All data preprocessing steps (scaling, PCA, and feature filtering) were executed exclusively within training partitions (inner or outer loops), then applied to validation or test partitions using stored transformation parameters. This practice prevents data leakage from test or validation sets into the preprocessing step, thereby preserving the integrity of model evaluation [42]. Multiple test-size splits (20%, 30%, 40%) to variation in train/test partition affects performance metrics. For each split, researchers conducted cross-validation with both 3-fold and 5-fold outer loops. The outer cross-validation was repeated multiple times with different random seeds to capture variability in performance estimates due to random data partitioning. This repetition enhances robustness by evaluating how sensitive the models are to different data splits. For each outer-fold test set, I recorded accuracy, macro-F1, ROC-AUC, and confusion matrices. Assess both overall performance and class-level behaviour, which is particularly important in the presence of class imbalance. Hyperparameter tuning was performed within the inner cross-validation loops. For each candidate model (e.g., SVM, Random Forest, Logistic Regression), researchers searched over a grid of hyperparameters (e.g., C and kernel for

SVM, regularisation strength for Logistic Regression, and the number of estimators for Random Forest). Researchers selected the hyperparameter set that yielded the best average performance (macro-F1 or ROC-AUC) on the inner validation folds. Then, researchers trained the model on the full outer training set with the selected hyperparameters and evaluated it on the outer test set.

3.6. Hyperparameter Tuning

The model “Integration Learning” adds value to concatenate standardised per-omic features for intersection samples → train Elastic-Net Logistic Regression and Linear SVM with class weights:

- **Logistic Regression:** C ∈ {5}; panelty=l1_ratio, solver='liblinear', max_iter=3000).
- **SVM:** kernel='linear', probability=True.
- **RF:** n_estimators=100, random_state=42.

Ensuring the reproducibility of experiments is a cornerstone of reliable, transparent results in computational studies involving machine learning. The reproducibility was maintained by ensuring that all steps in the data preprocessing pipeline, model training, and evaluation were fully transparent:

- **Controlled Data Preprocessing:** Feature scaling and PCA were performed only on the training partitions to prevent data leakage, ensuring that no test-set information influenced model training. The separation between training and testing data is critical in maintaining valid performance estimates and preventing overestimation of model accuracy.
- **Random Seed:** To achieve accurate results, random seeds were fixed across all processes involving randomisation, including cross-validation splits and model training, to ensure that each experiment run produces consistent results regardless of external factors such as hardware or software variations. Constant random seeds would help replicate the precise conditions and the outcomes postulated by the code for any given experimenter.
- **Model and Data Splits:** Nested cross-validation was used to compare the model's performance across various data splits. Not only does methodology help enable consistent assessment, but it also helps reduce the risk of overfitting, as it offers a stronger defence against model performance across a wide range of data subsets. The reliability of the results is also enhanced by using multiple train-test splits (e.g., 60/40%, 70/30%, and 80/20%).
- **Hyperparameter Tuning:** Hyperparameter tuning optimised the models to ensure they performed at optimal levels while avoiding overfitting. For models such as Logistic Regression, Support Vector Machine (SVM), and Random Forest (RF), hyperparameters such as regularisation strength (C for LR), kernel type (linear for SVM), and the number of estimators (n_estimators) for RF were selected. Grid search and cross-validation were used to determine the optimal parameter settings for each model. Regularisation and model complexity (for SVM and RF) influence performance, and tuning these parameters improved generalisation to unseen data.
- **Reproducibility of Model and Results:** All model configurations, including hyperparameters, training setup, and preprocessing steps, were stored and logged to ensure full reproducibility:
    - Model hyperparameters (C, max_iter, kernel)
    - Preprocessing (scaling, PCA transformations)
    - Software environment used (versions of Python, scikit-learn, pandas)
    - Training and validation data splits

These principles ensure that the study's results are not only valid and consistent but also independently verifiable, as the code, data, and documentation for each experiment should be publicly available.

3.6.1. Risks and Mitigations

- **High-p / Low-n Overfitting:** Imbalanced classes verified by the CV method.
- **Class Imbalance:** Class-weighted loss and macro metrics.
- **Compute/Time Limits:** Heavier deep models are intentionally scoped out.

4. Implementation and Experiments

In this paper, the machine learning models were implemented, and the experimental setup for evaluating performance in predicting cancer subtypes was described. Data processing, model training, cross-validation, and the tools and libraries used throughout the experiments.

4.1. Experimental Setup

The experiments were carried out in a Python environment, using popular libraries such as pandas, numpy, scikit-learn, and matplotlib for data manipulation, model training, evaluation, and visualisation. The experiments were run on a local computer using Python 3.8, and all required dependencies were managed with Anaconda to ensure consistent results across runs. The model training and testing datasets were obtained from The Cancer Genome Atlas (TCGA) for LGG and BRCA cancer types. This dataset has several subtypes according to clinical definitions; however, in this analysis, only gene expression data (transcriptomic matrices) were considered, as other omics data were not included.

4.2. Data Preparation and Preprocessing

4.2.1. Data Import and Cleaning

The gene expression results were converted to a pandas DataFrame. Attributes (genes) with zero variance across all samples were dropped, as they provided no useful information for classification. I also ensured that sample identifiers were properly matched and corrected any errors to maintain the dataset's accuracy.

4.2.2. Handling Class Imbalance

Because the BRCA and LGG datasets were unbalanced (some subtypes had much fewer samples), I employed stratified sampling when splitting the training and test sets. This ensured consistent class representation across splits, helping mitigate bias and enhancing the model's fairness in training and evaluation.

4.2.3. Standardisation and PCA

I normalised the features by Z-scoring them to a common scale. I used Principal Component Analysis (PCA) to reduce the data size thereafter. I chose the number of principal components to keep at least 95 per cent of the original variance, reduce noise in the analysis, and make the analysis more effective.

4.2.4. Splitting the Data

Data was split into training and test sets to achieve different ratios (20 per cent, 30 per cent, and 40 per cent). This helped me determine the effect of the training set size on the model's performance. I performed cross-validation (3- and 5-fold) after each split to assess the strength and generalisation of the models. I used nested cross-validation to prevent overfitting and, at the same time, ensure that hyperparameter tuning was not sensitive to model evaluation.

4.3. Model Selection and Training

4.3.1. Baseline Models

Three baseline models were for cancer subtype classification:

- **Logistic Regression:** A type of linear model that classifies data based on learned probabilities.
- **Support Vector Machine:** It is a linear kernel classification model that is well-suited to high-dimensional data.
- **Random Forest:** It is an ensemble technique under which several decision trees are created to make classification of data, and therefore, it is resistant to overfitting in small datasets.

For each baseline model, PCA was used as a preprocessing step to reduce data dimensionality. The PCA transformation was fitted on the training data in each cross-validation fold, and the same transformation was applied to the test set to ensure. Hyperparameters for each model were tuned using grid search combined with cross-validation. For Logistic Regression, the regularisation parameter (C) was tuned; for SVM, both the regularisation parameter (C) and the kernel type were optimised. For Random Forest, the number of trees (n_estimators) and maximum depth (max_depth). The best-performing hyperparameters, based on macro-F1 score and ROC-AUC, were used for final model training.

4.4. Evaluation Protocol

Some of the key metrics that were used to determine model performance were:

- **Accuracy:** This is the number of correct predictions made as compared to the number of predictions made.

- **Macro-F1 Score:** Since the dataset had an unequal class distribution, I utilised the macro-F1 score to ensure that all classes, regardless of their frequency, were given equal weight in the model's performance.
- **ROC-AUC:** Area Under the Curve (ROC-AUC) was used to determine how effectively the model classifies the classes.
- **Confusion Matrices:** They were constructed to visualise each model's performance and are analysed at the subtype level.

The cross-validation procedure was better structured, preventing any data leakage. The only preprocessing procedures, namely scaling and PCA, were applied to the training sets at every iteration, but not to the test data, so the model was not fit to the test data.

4.5. Experiment Reproducibility

To repeat the experiments, the researcher fixed the random seeds for all data splits, hyperparameter tuning, and model training. This ensured that every experiment was carried out under the same conditions. All preprocessed data, trained models, PCA components, and scaling parameters were stored and registered, allowing the experiments to be rerun. The experiments were conducted within a controlled Python environment, and the entire setup was well documented, in keeping with the standards of reproducible research.

4.6. Tooling and Libraries Used

The experiments were coded in Python, and they used the following libraries:

- **Pandas:** The application of statistical operations.
- **Numpy:** Numerical models and array models.
- **Scikit-Learn:** To adopt the machine learning models, PCA, and cross-validation.
- **Matplotlib and Seaborn:** Data visualisation, e.g., confusion matrices, ROC curves, and performance metrics.

The experiments involve simple machine learning models for predicting cancer subtypes using gene expression as input, with PCA for dimensionality reduction and an extensive cross-validation strategy to achieve credible results. It was conducted to determine the influence of PCA on model performance, class imbalance, and interpretability. These experiments will be presented in the following paper and will report on each model's performance and the information obtained from the analysis.

5. Results and Discussion

5.1. Descriptive Analysis and Data Cleaning

The raw datasets for LGG and BRCA were first imported into Python. The unwanted variables, such as patient_id, OS, OS_days, clinical_ajcc_pathologic_tumor_stage, and clinical_clinical_stage, were excluded from this study. The empty columns from the 16375 input features of the LGG and BRCA dataset were removed. Then, all 16375 features from both the BRCA and LGG datasets were standardised to improve prediction accuracy. The frequency distributions and descriptive statistics, including minimum, maximum, counts, average, and standard deviation, for the demographic variables in both the LGG and BRCA datasets were estimated. There were three types of LGG dataset: 66 Females had Astrocytoma, 50 had Oligoastrocytoma, and 68 had Oligodendroglioma. Similarly, 77 males had Astrocytoma, 60 had Oligoastrocytoma, and 91 had Oligodendroglioma. Most males had LGG disorders rather than females. The minimum age of the patients was 14 years, the maximum was 87 years, with an average age of 42.92 years and a 13.36-year age gap among the patients for the LGG data given in Table 1. Similarly, there were eight types of BRCA dataset, where 1 female had Infiltrating Carcinoma NOS, 573 had Infiltrating Ductal Carcinoma, and 114 had Infiltrating Lobular Carcinoma, 5 had Medullary Carcinoma, 3 had Metaplastic Carcinoma, 21 females had Mixed Histology, 11 had Mucinous Carcinoma, and 34 had other specified BRCA type disease. Similarly, 8 males had Infiltrating Ductal Carcinoma, and 1 had other specified BRCA type disease. More males had LGG disorders than females.

Table 1: Descriptive statistics of LGG data with gender and age

Disease Type	Gender		Total	Measures	Age
	Female	**Male**		Minimum	14
Astrocytoma	66	77	143	Mean	42.92
Oligoastrocytoma	50	60	110	Maximum	87

Oligodendroglioma	68	91	159	Standard deviation	13.36
Total	184	228	412	Counts	412

The minimum age of the BRCA patients was 26 years, the maximum was 90 years, with an average age of 57.85 years and a 13.29-year age gap among the patients, as shown in Table 2.

Table 2: Descriptive statistics of BRCA data with gender and age

Disease Type	**Gender**		**Total**	**Measures**	**Age**
	Female	**Male**		Minimum	26
Infiltrating Carcinoma NOS	1		1	Mean	57.85
Infiltrating Ductal Carcinoma	573	8	581	Maximum	90
Infiltrating Lobular Carcinoma	114		114	Standard deviation	13.29
Medullary Carcinoma	5		5	Counts	771
Metaplastic Carcinoma	3		3		
Mixed Histology	21		21		
Mucinous Carcinoma	11		11		
Other, specify	34	1	35		
Total	762	9	771		

5.2. LGG Data Analysis

The empty columns in the LGG datasets were removed to improve data robustness and classification performance.

5.2.1. Risk Identification

The risk of LGG subtypes by age is shown in Table 3. The age variable was divided into three groups (<40, 40-59, and 60+) to assess the risk of Astrocytoma, Oligoastrocytoma, and Oligodendroglioma among 412 patients.

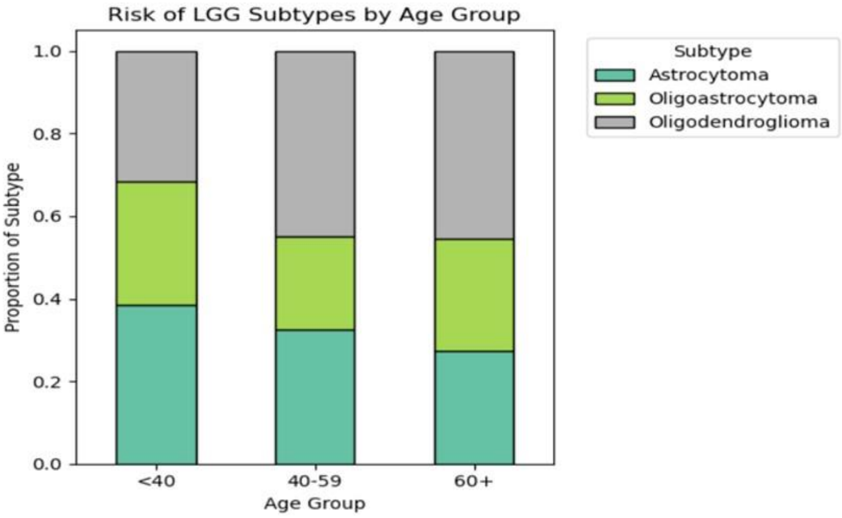


Figure 2: Risk of LGG subtype with respect to age group

The cross-tabulation of risks, with a chi-square value, shows that age and LGG subtypes are significantly related. The results revealed that the highest proportion of Astrocytoma (38.6%) was observed among patients aged <40 years. Similarly, the highest proportion of Oligoastrocytoma (45%) was found among 40-59-year-old patients.

Table 3: Risk identification of LGG subtypes by age group

LGG type/Age	**Astrocytoma**	**Oligoastrocytoma**	**Oligodendroglioma**	**Total**	**Chi-Square**
<40	76/197= (38.6%)	59/197 = (29.9%)	62/197 = (31.5%)	197	0.001

40-59	52/160 = (32.5%)	36/160 = (22.5%)	72/160 = (45.0%)	160	
60+	15/55 = (27.3%)	15/55 = (27.3%)	25/55 = (45.5%)	55	

Similarly, the highest proportion of Oligodendroglioma (45%) was found among patients aged 60+. The graphical presentation of the risk of the LGG subtype by age group is shown in Figure 2.

5.2.2. Baseline Models with Cross-Validation

The baseline models were applied to different training/test partition sizes, such as 80%/20%, 70%/30%, and 60%/40%. The accuracy rates, F1 macro, and ROC AUC for logistic regression (LR), SVM, and random forest (RF) on the training and test datasets are shown in Table 4. All baseline models achieved 100% accuracy (F1-MARCO) and ROC-AUC on the training datasets across the 80%/20%, 70%/30%, and 60%/40% data splits. But the baseline models achieved very low test accuracy. Like the LR model achieved 56.60%, 52.40%, and 51.80% accuracy rates for each 80%/20%, 70%/30%, and 60%/40% data sizes respectively. Similarly, the SVM model achieved 51.80%, 50%, and 55.10% accuracy rates for each 80%/20%, 70%/30%, and 60%/40% data sizes respectively. Similarly, the RF model achieved 60.20%, 58%, and 56.90% accuracy rates for each 80%/20%, 70%/30%, and 60%/40% data sizes respectively. Overall, the baseline model did not achieve high accuracy on the test datasets with the proposed parameters and also exhibited overfitting.

Table 4: Classification report of baseline models of training and testing data

Dataset Partition	Partition	Logistic Regression				SVM		Random Forest		
		20%	30%	40%	20%	30%	40%	20%	30%	40%
Training	Accuracy	1	1	1	1	1	1	1	1	1
	F1-macro	1	1	1	1	1	1	1	1	1
	ROC-AUC	1	1	1	1	1	1	1	1	1
Testing	Accuracy	0.566	0.524	0.527	0.518	0.5	0.551	0.602	0.580	0.569
	F1-macro	0.512	0.490	0.496	0.478	0.465	0.529	0.529	0.515	0.523
	ROC-AUC	0.717	0.688	0.671	0.710	0.693	0.698	0.739	0.732	0.719

The baseline models were also assessed using 3- and 5-fold cross-validation. The cross-validation results are given in Table 4. The LR, SVM, and RF models achieved average accuracies of 51.95%, 54.86%, and 56.80%, respectively, in 3-fold cross-validation. Similarly, the LR, SVM, and RF models achieved average accuracies of 51.95%, 55.32%, and 58.72%, respectively, across 5-fold cross-validation (Table 5).

Table 5: Cross-validation report of baseline models based on testing data

Number of Folds		Logistic Regression		SVM		Random Forest	
		Average Accuracy	Average f1-marco	Average Accuracy	Average f1-marco	Average Accuracy	Average f1-marco
Folds	3	51.95%	0.4725	54.86%	0.5191	56.80%	0.504
	5	51.95%	0.466	55.32%	0.5257	58.72%	0.5165

5.3. Integrated Models

The integrated models were applied to different training/test partition sizes, such as 80%/20%, 70%/30%, and 60%/40%. The accuracy rates, f1-macro, and ROC-AUC for LR+PCA, SVM+PCA, and RF+PCA on the training and test datasets are shown in Table 6. The LR+PCA model achieved 71.70%, 72.60%, and 78.10% accuracies on the training data and 53%, 50.80%, and 55.20% accuracies on the testing data (80%/20%, 70%/30%, and 60%/40%), respectively. Similarly, the SVM+PCA model achieved 82.10%, 82.20%, and 87.90% accuracies on the training data and 57.80%, 52.40%, and 55.20% accuracies on the testing data (80%/20%, 70%/30%, and 60%/40%), respectively.

Table 6: Classification report of integrated models of training and testing data

Dataset Partition	Partition	Logistic Regression + PCA				SVM+PCA		Random Forest + PCA		
		20%	30%	40%	20%	30%	40%	20%	30%	40%
Training	Accuracy	0.717	0.726	0.781	0.821	0.872	0.879	1	1	1
	F1-macro	0.693	0.709	0.762	0.814	0.863	0.866	1	1	1
	ROC-AUC	0.881	0.893	0.911	0.836	0.871	0.838	1	1	1

Testing	Accuracy	0.530	0.508	0.552	0.578	0.524	0.552	0.590	0.540	0.546
	F1-macro	0.494	0.470	0.519	0.541	0.500	0.533	0.517	0.489	0.500
	ROC-AUC	0.715	0.703	0.712	0.691	0.651	0.669	0.759	0.686	0.706

But the RF+PCA model achieved 100%, 100%, and 100% accuracies on training data, and 59%, 54%, and 54.60% accuracionn testingdata data (80%/20%, 70%/30%, and 60% Overall, the integrated models achieved higher test accuracy than the testing models, using the proposed parameters. The integrated models also reduce overfitting. The integrated models were also assessed using 3- and 5-fold cross-validation. The cross-validation results are given in Table 6. The LR+PCA, SVM+PCA, and RF+PCA models achieved average accuracies of 53.65%, 50.73%, and 57.04%, respectively, across 3-fold cross-validation. Similarly, the LR, SVM, and RF models achieved average accuracies of 56.32%, 54.11%, and 50.43%, respectively, across 5-fold cross-validation (Table 7).

Table 7: Cross-validation report of integrated models based on testing data

Number of Folds		Logistic Regression + PCA		SVM+PCA		Random Forest + PCA	
		Average Accuracy	Average f1-marco	Average Accuracy	Average f1-marco	Average Accuracy	Average f1-marco
Folds	3	53.65%	0.4995	50.73%	0.5234	57.04%	0.5043
	5	56.32%	0.5336	54.11%	0.4880	50.43%	0.5223

5.4. Confusion Matrix Comparison

Both baselines and integrated models for the LGG dataset were also compared using confusion matrices, where the diagonal entries indicate the patients' correct predictions for the three LGG disease types.

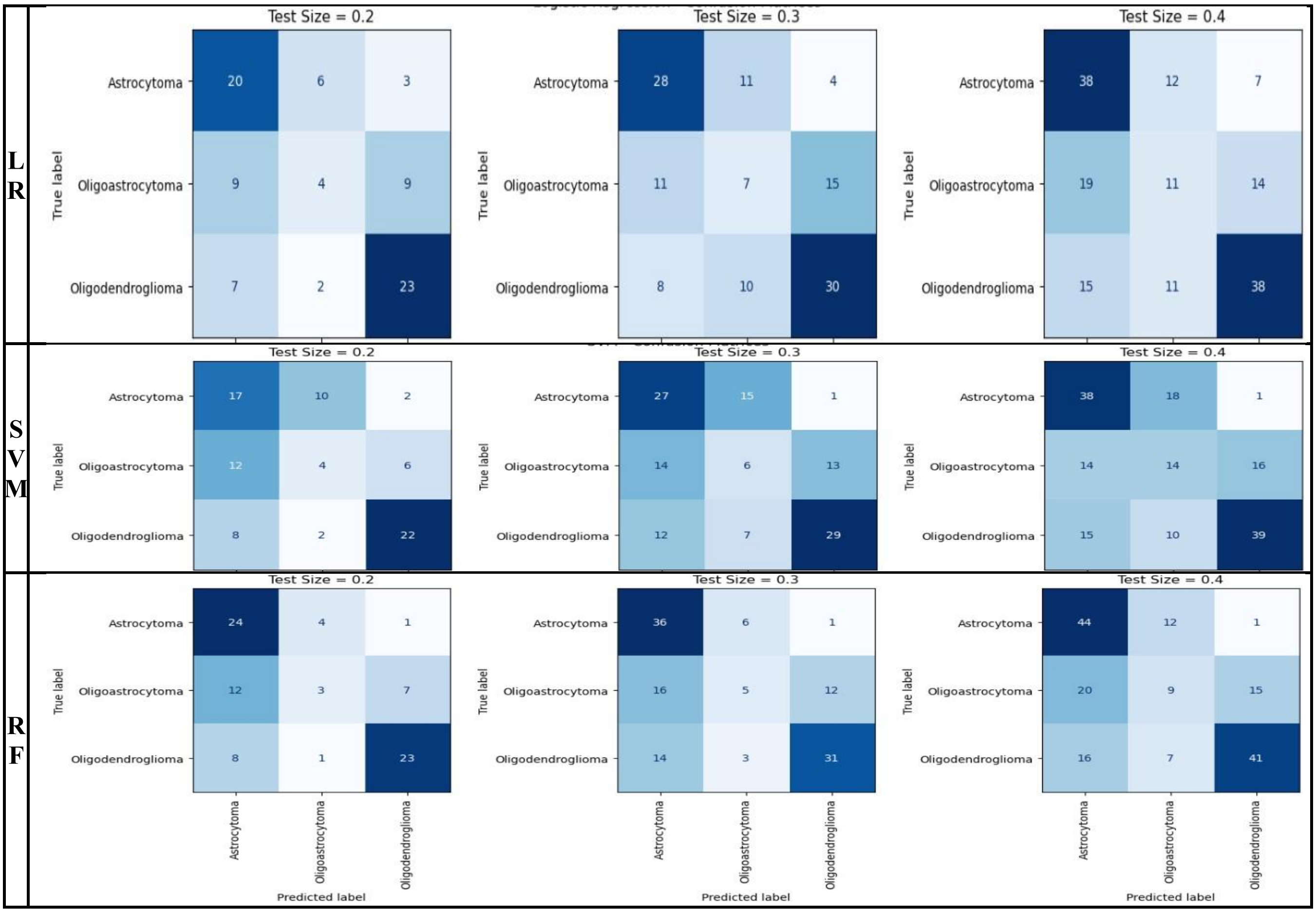


Figure 3: Confusion matrices of baseline models with respect to different data sizes

Again, the integrated models achieved higher prediction rates than the baseline models for LGG subtypes detection. The detailed confusion matrices for both the baseline and integrated models are presented in Figures 3 and 4.

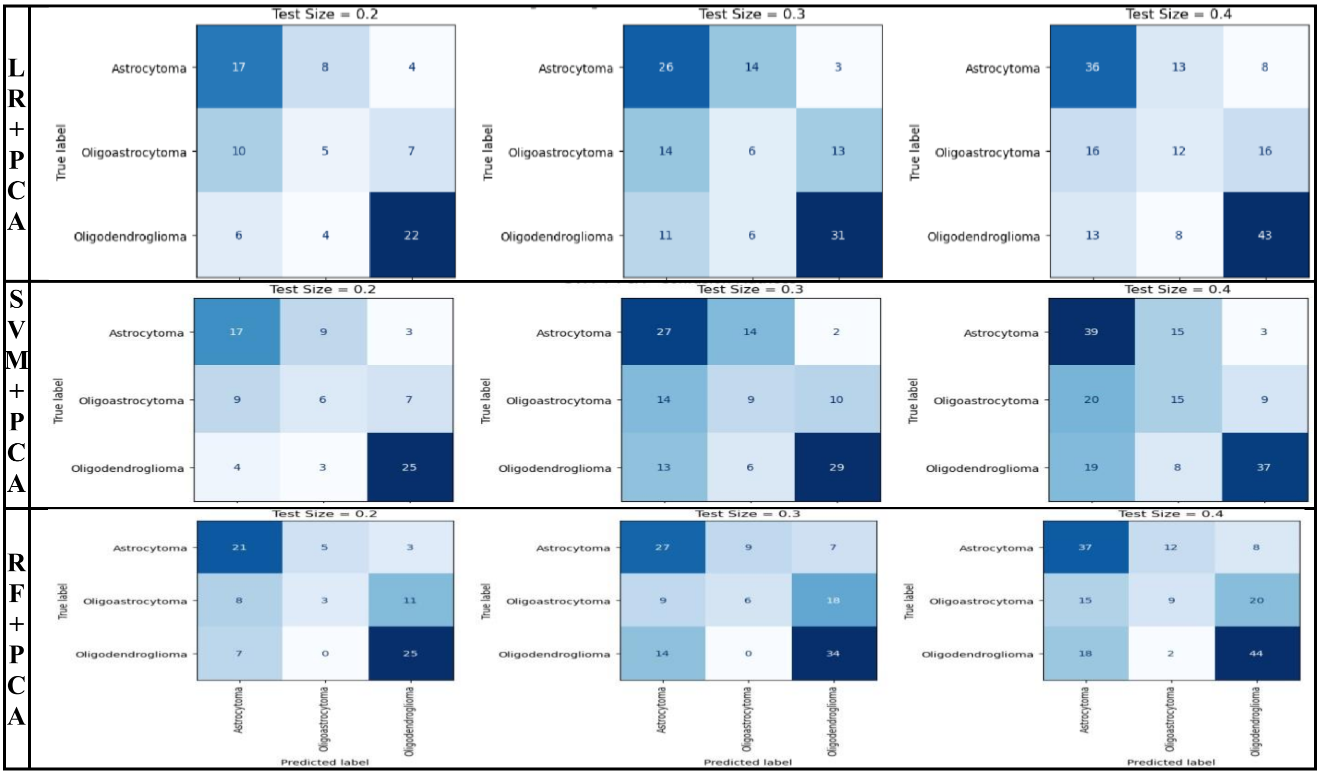


Figure 4: Confusion matrices of integrated models with respect to different data sizes

5.5. BRCA Data Analysis

All male patients were excluded from the BRCA dataset; only include Infiltrating Ductal Carcinoma and Infiltrating Lobular Carcinoma BRCA types. All others were classified as other specified BRCA disease due to class imbalance in the classifier models. The empty columns in the BRCA datasets were removed to improve data robustness and classification performance.

5.5.1. Risk Identification

The risk of the BRCA subtypes by age is shown in Table 8. The age variable was converted into three classes (<40, 40-59, and 60+) to assess the risk of Infiltrating Ductal Carcinoma, Infiltrating Lobular Carcinoma, and Other. This analysis included 672 patients. The cross-tabulation of risks, with a chi-square value, indicates that age and BRCA subtypes are significantly related. The results revealed that the highest proportion of Infiltrating Ductal Carcinoma (90%) was found among patients <40 years old. Similarly, the highest rate of Infiltrating Ductal Carcinoma (76%) was found among patients aged 40-59 years.

Table 8: Risk identification of BRCA subtypes by age group

BRCA type/Age	Infiltrating Carcinoma	Ductal	Infiltrating Carcinoma	Lobular	Other, specify	Chi-Square
<40	54/60 = 90.0%		0/60 = 0.0%		6/60 = 10.0%	0.001
40-59	275/362 = 76.0%		50/362 = 13.8%		37/362 = 10.2%	
60+	244/340 = 71.8%		64/340 = 18.8%		32/340 = 9.4%	

Similarly, the highest proportion of Infiltrating Ductal Carcinoma (71.8%) was found among patients aged 60+. The Infiltrating Ductal Carcinoma breast cancer type has a higher risk in each age group of patients than other types. The graphical presentation of the risk of BRCA subtypes by age group is shown in Figure 5.

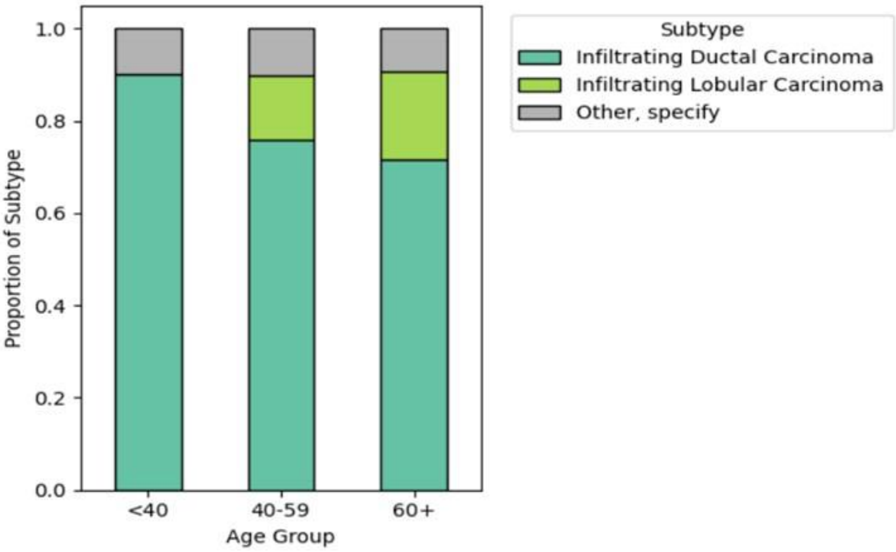


Figure 5: Risk of BRCA subtype with respect to age group

5.5.2. Baseline Models

The baseline models were applied to different training/test partition sizes, such as 80%/20%, 70%/30%, and 60%/40%. The accuracy rates, f1-macro, and ROC-AUC for LR, SVM, and RF on the training and test datasets are shown in Table 9. All baseline models achieved 100% accuracy (F1-MARCO) and ROC-AUC on the training datasets across the 80%/20%, 70%/30%, and 60%/40% data splits. But the baseline models achieved low test accuracy. Like the LR model achieved 83%, 78.10%, and 80% accuracy rates for each 80%/20%, 70%/30%, and 60%/40% data sizes respectively. Similarly, the SVM model achieved 81.70%, 82.10%, and 80.30% accuracy rates for each 80%/20%, 70%/30%, and 60%/40% data sizes respectively. Similarly, the RF model achieved 77.70%, 75.90%, and 78.80% accuracy rates for each 80%/20%, 70%/30%, and 60%/40% data sizes respectively. Overall, the baseline model achieved good accuracy on the test datasets with the proposed parameters, but it also exhibited overfitting (Table 9).

Table 9: Classification report of baseline models of training and testing data

Dataset Partition	**Partition**	**Logistic Regression**				**SVM**		**Random Forest**		
		20%	**30%**	**40%**	**20%**	**30%**	**40%**	**20%**	**30%**	**40%**
Training	Accuracy	1	1	1	1	1	1	1	1	1
	F1-macro	1	1	1	1	1	1	1	1	1
	ROC-AUC	1	1	1	1	1	1	1	1	1
Testing	Accuracy	0.830	0.781	0.800	0.817	0.821	0.803	0.777	0.759	0.780
	F1-macro	0.611	0.554	0.585	0.622	0.622	0.576	0.389	0.387	0.426
	ROC-AUC	0.785	0.791	0.782	0.744	0.744	0.761	0.821	0.747	0.804

The baseline models were also assessed using a 3- and 5-fold cross-validation method on the BRCA dataset. The cross-validation results are given in Table 8. The LR, SVM, and RF models achieved average accuracies of 81.50%, 78.74%, and 77.69%, respectively, using a 3-fold cross-validation method. Similarly, the LR, SVM, and RF models achieved average accuracies of 82.03%, 80.19%, and 77.95%, respectively, across 5-fold cross-validation.

Table 10: Cross-validation report of baseline models based on testing data

Number of Folds		**Logistic Regression**		**SVM**		**Random Forest**	
		Average Accuracy	**Average f1-Marco**	**Average Accuracy**	**Average f1-Marco**	**Average Accuracy**	**Average f1-Marco**
Folds	3	81.50%	0.8203	78.74%	0.5448	77.69%	0.4196
	5	82.03%	0.6081	80.19%	0.5749	77.95%	0.5275

5.5.3. Integrated Models

The integrated models were applied to different training/test partition sizes, such as 80%/20%, 70%/30%, and 60%/40%. The accuracy rates, f1-macro, and ROC-AUC for LR+PCA, SVM+PCA, and RF+PCA on the training and test datasets are shown in Table 10. The LR+PCA model achieved 86%, 87.40%, and 87.50% accuracies from training, and 81%, 79.40%, and 80.03% accuracies from testing (80%/20%, 70%/30%, and 60%/40%) data sizes, respectively. Similarly, the SVM+PCA model achieved 87.50%, 89.10%, and 88.80% accuracies on the training data and 80.30%, 78.60%, and 79.30% accuracies on the testing data (80%/20%, 70%/30%, and 60%/40%), respectively. But the RF+PCA model achieved 100%, 100%, and 100% accuracies on the training data and 77.10%, 77.20%, and 76.30% accuracies on the test data (80%/20%, 70%/30%, and 60%/40%), respectively. Overall, the integrated models achieved higher accuracy on the test datasets than the baseline models, using the proposed parameters. The integrated models also reduce overfitting.

Table 11: Classification report of integrated models of training and testing data of BRCA

Dataset Partition	Partition	Logistic Regression+PCA				SVM+PCA		Random Forest+PCA		
		20%	30%	40%	20%	30%	40%	20%	30%	40%
Training	Accuracy	0.860	0.874	0.864	0.875	0.891	0.888	1	1	1
	F1-macro	0.693	0.706	0.685	0.688	0.718	0.717	1	1	1
	ROC-AUC	0.882	0.881	0.879	0.865	0.850	0.859	1	1	1
Testing	Accuracy	0.810	0.794	0.803	0.803	0.786	0.793	0.771	0.772	0.763
	F1-macro	0.569	0.546	0.550	0.49 4	0.519	0.533	0.404	0.387	0.369
	ROC-AUC	0.766	0.758	0.791	0.774	0.756	0.769	0.830	0.732	0.762

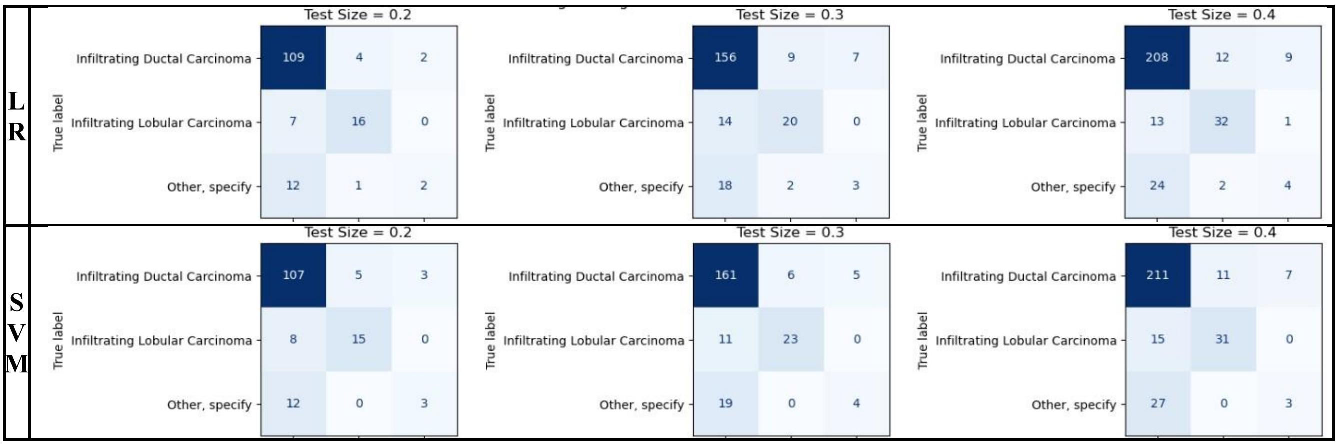
The integrated models were also assessed using 3- and 5-fold cross-validation. The cross-validation results are given in Table 11. The LR+PCA, SVM+PCA, and RF+PCA models achieved average accuracies of 80.84%, 80.58%, and 76.64%, respectively, using a 3-fold cross-validation method. Similarly, the LR, SVM, and RF models achieved average accuracies of 80.58%, 80.05%, and 77.69%, respectively, across 5-fold cross-validation (Table 12).

Table 12: Cross-validation report of integrated models based on testing data of BRCA

Number of Folds		Logistic Regression+PCA		SVM+PCA		Random Forest+PCA	
		Average Accuracy	Average f1-marco	Average Accuracy	Average f1-marco	Average Accuracy	Average f1-marco
Folds	3	80.84%	0.5618	80.58%	0.5482	76.64%	0.3682
	5	80.58%	0.5574	80.05%	0.5213	77.69%	0.3886

5.5.4. Confusion Matrix Comparison

Both baselines and integrated models for the BRCA dataset were also compared using confusion matrices, where the diagonal entries indicate the number of patients correctly predicted for the three BRCA disease types.



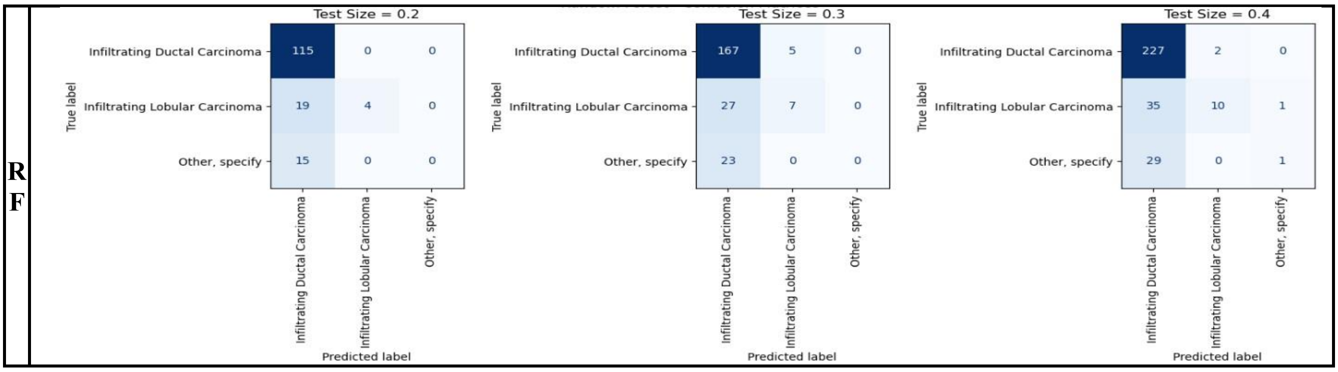


Figure 6: Confusion matrices of baseline models with respect to different data sizes

Again, the integrated models achieved higher prediction rates than the baseline models for BRCA subtypes detection. The detailed confusion matrices for both the baseline and integrated models are presented in Figures 6 and 7.

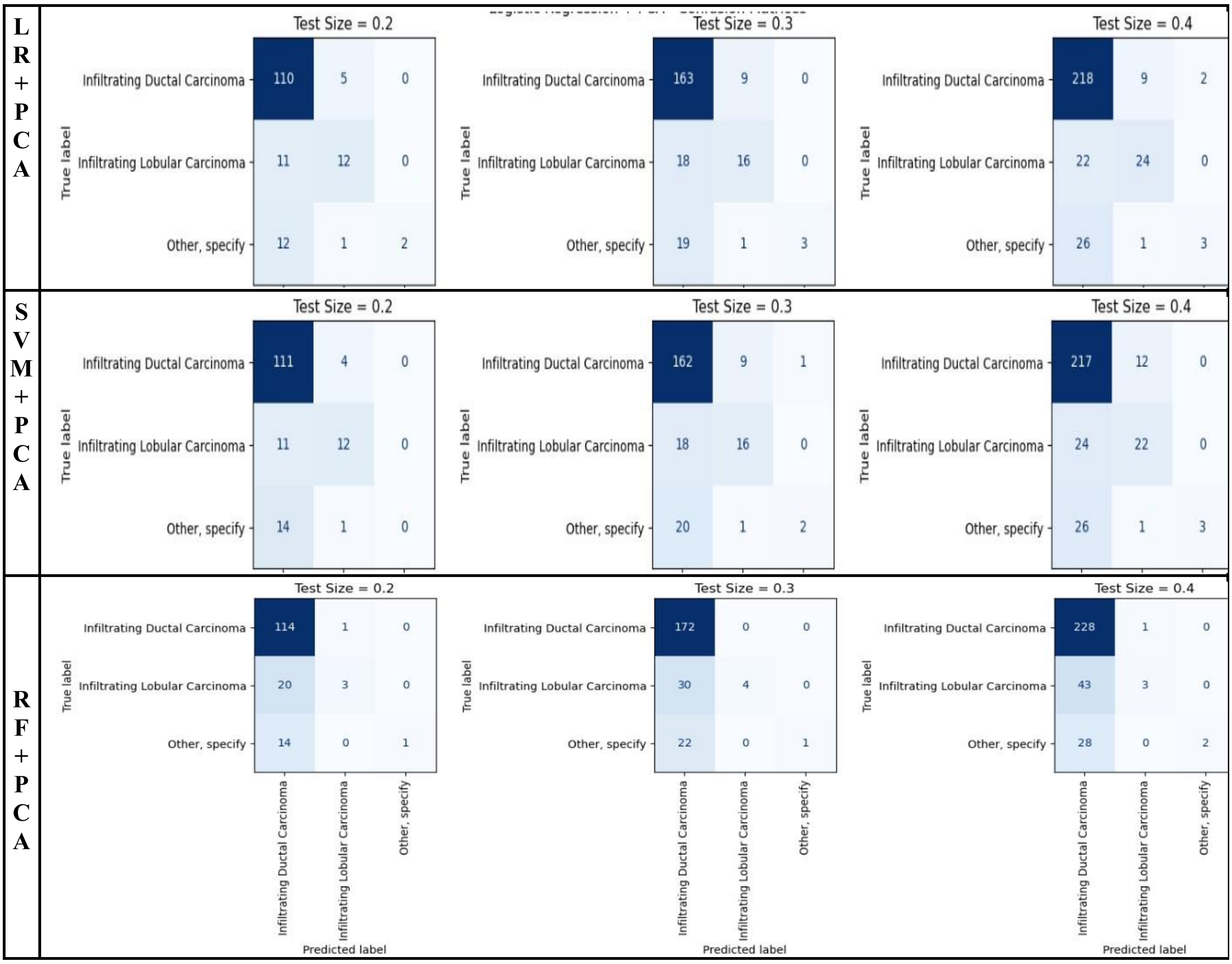


Figure 7: Confusion matrices of integrated models with respect to different data sizes

5.6. Comparison with Literature and Class Imbalance Considerations

In studies of cancer subtype classification using gene-expression data, researchers often face high dimensionality and class imbalance. A recent review on machine learning for cancer classification showed that classical machine-learning methods remain competitive under such constraints [13]. For instance, a 2024 study on breast cancer gene-expression data applied resampling methods and demonstrated improved performance of support vector machines and random forests when class imbalance was addressed [18]. Another study integrating omics data (RNA-seq, copy number variation, methylation) compared 18 classifiers and tested several resampling strategies. Their results indicated that the Synthetic Minority Over-sampling Technique often yielded the best balanced performance (accuracy, AUC, recall) across underrepresented classes [48]. More recent work used explainable-ML techniques (SHAP, accumulated local effects) after feature selection to ensure both high classification accuracy and interpretability of gene contributions [17]. Compared with these studies, the work uses dimensionality reduction (PCA) before classification. This reduces the feature space while preserving most of the variance, helping mitigate overfitting and instability. Class-imbalance handling through class-weighting (or alternate techniques) parallels the resampling or cost-sensitive strategies seen elsewhere. However, many papers that report high accuracy also use explicit resampling or hybrid feature-selection + model-based resampling + explainable ML pipelines. That suggests a potential gap in approach: using PCA alone may not fully address imbalance effects on minority classes.

6. Conclusion

Using gene-expression data from LGG and BRCA datasets, the paper has critically evaluated the utility of classical machine-learning regressions based on Principal Component Analysis (PCA) for classifying cancer subtypes. The main goal was to explore dimensionality reduction using PCA to enhance the quality, strength, and understanding of the models, especially given the constraints of high-dimensional genomic information and limited sample sizes. PA can potentially alleviate the problem of overfitting, and the impact on increasing model performance in a complex environment had to be carefully scrutinised. The study hypothesis was an independent hypothesis testing the claim that dimensional reduction would yield more generalizable models. However, once more, this also raises the question of whether PCA is adequate for addressing these underlying issues in genomic data, at least where the biological implications of the principal components are less evident. The study focused on predicting cancer subtypes from gene expression data, and the models discussed included Logistic Regression, SVM, and Random Forest. The models also performed better in generalisation when coupled with PCA, with the highest performance rate recorded in the SVM+PCA model, which is the most accurate on the LGG and BRCA datasets. The results of the prediction process indicated that SVM+PCA was a superior model compared to the reference models, achieving significant gains in accuracy, macro-F1, and ROC-AUC. The metrics were used to determine the classification threshold and to achieve balanced performance across subtypes, including those with fewer samples. This confirmed that subtype prediction in cancer was effective and that PCA was necessary to improve the model's performance and address the high-dimensional data problem.

6.1. Model Performance

The findings showed that the baseline models (Logistic Regression, Support Vector Machines, and Random Forests) exhibited some overfitting to the training data, but adding PCA improved model generalisation. The models perform well on unseen test data when PCA is used. The best test accuracy on the LGG and BRCA datasets is achieved with SVM+PCA, compared to the other models. In particular, SVM+PCA showed increases in macro-F1, ROC-AUC, and accuracy, suggesting that dimensionality reduction reduced overfitting and decreased noise in the high-dimensional feature space. This emphasises the effectiveness of PCA in enhancing model robustness. However, it also raises questions about finding the right balance between dimension reduction and preserving sufficient biological signal in the data.

6.2. Class Imbalance Handling

One of the biggest problems in this study was class imbalance, especially in the BRCA dataset. This was solved, however, through cross-validation and stratification, in which a sample population was split into folds with the same proportions of each cancer subtype. This fact was reinforced by the use of macro-F1 and ROC-AUC metrics, which are better suited to handling imbalanced datasets. Conversely, using accuracy alone would have been deceptive because of the unequal class distribution, which should be taken into account when assessing model performance.

6.3. Reproducibility and Cross-Validation

Nested cross-validation was employed to prevent optimistic bias, and the processes of model tuning and performance evaluation were carried out independently. The models were evaluated using a series of test-size splits (20%, 30%, and 40%) to assess their stability across different splits. Multiple nested cross-validations with different random seeds yielded consistent results, indicating the stability of the models and the reproducibility of the results. It has adhered to best practices to achieve a leakage-free pipeline: all preprocessing (scaling, PCA, feature filtering) was performed within the training folds. It was a strategy to ensure an unbiased assessment and reduce the risk of data contamination during model training.

6.4. Interpretability

Another aspect highlighted in the study is the trade-offs in model interpretability when using PCA. Although PCA effectively reduced dimensionality and improved generalisation, it led to interpretability problems because the principal components are linear combinations of features that are not necessarily biologically meaningful. Nevertheless, assessing component loadings and explained variance enabled us to relate the most significant components to genes of interest in cancer biology. The results of this analysis provided interesting insights into the biological mechanisms that influence cancer subtype differentiation. They helped clarify the factors that affect model predictions, even after data transformation via PCA.

6.5. Limitations and Future Work

Although PCA integration improved model performance, the study has weaknesses because it used single-omics gene expression data. Future studies may improve this method by incorporating multi-omics data (e.g., genomics, methylation, and proteomics), which may offer greater predictive power and deeper insight into the cellular processes underlying cancer. Computational resources were also limited for using more complex models, such as deep learning, in the paper. High-performance computing could provide useful insights, and its predictive capabilities could be enhanced for cancer subtypes.

Acknowledgement: The authors gratefully acknowledge the academic support and facilities provided by Northumbria University, Newcastle upon Tyne and the University of the West of Scotland.

Data Availability Statement: The study uses a dataset focused on improving cancer subtype classification using high-dimensional gene expression data via PCA and machine learning.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest Statement: The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Ethics and Consent Statement: Ethical approval was obtained prior to data collection, and informed consent was secured from all participating organizations and individuals.

References

1. A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, “Benchmark of filter methods for feature selection in high-dimensional gene expression survival data,” *Briefings in Bioinformatics*, vol. 23, no. 1, pp. 1–13, 2021.
2. A. R. Baião, Z. Cai, R. C. Poulos, P. J. Robinson, R. R. Reddel, Q. Zhong, S. Vinga, and E. Gonçalves, “A technical review of multi-omics data integration methods: From classical statistical to deep generative approaches,” *Briefings in Bioinformatics*, vol. 26, no. 4, pp. 1–18, 2025.
3. A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K. A. Lê Cao, “DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays,” *Bioinformatics*, vol. 35, no. 17, pp. 3055–3062, 2019.
4. A. Statnikov, L. Wang, and C. F. Aliferis, “A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification,” *BMC Bioinformatics*, vol. 9, no. 1, pp. 1–10, 2008.
5. B. M. S. Hasan and A. M. Abdulazeez, “A review of principal component analysis algorithm for dimensionality reduction,” *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 20–30, 2021.
6. B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, “Similarity network fusion for aggregating data types on a genomic scale,” *Nature Methods*, vol. 11, no. 3, pp. 333–337, 2014.
7. C. Ambroise and G. J. McLachlan, “Selection bias in gene extraction based on microarray gene-expression data,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6562–6566, 2002.
8. D. Acharya and A. Mukhopadhyay, “A comprehensive review of machine learning techniques for multi-omics data integration: Challenges and applications in precision oncology,” *Briefings in Functional Genomics*, vol. 23, no. 5, pp. 549–560, 2024.
9. D. Sugianto and T. Wahyuningsih, “Classifying vehicle categories based on technical specifications using random forest and SMOTE for data augmentation,” *International Journal for Applied Information Management*, vol. 5, no. 4, pp. 179–191, 2025.

10. E. Pellegrino, C. Jacques, N. Beaufils, I. Nanni, A. Carlioz, P. Metellus, and L. Ouafik, “Machine learning random forest for predicting oncosomatic variant NGS analysis,” *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.
11. E. Withnell, X. Zhang, K. Sun, and Y. Guo, “XOmiVAE: An interpretable deep learning model for cancer classification using high-dimensional omics data,” *Briefings in Bioinformatics*, vol. 22, no. 6, pp. 1–11, 2021.
12. F. A. Ghaleb, F. Saeed, M. Al-Sarem, S. N. Qasem, and T. A. Hadhrami, “Ensemble synthesized minority oversampling-based generative adversarial networks and random forest algorithm for credit card fraud detection,” *IEEE Access*, vol. 11, no. 8, pp. 89694–89710, 2023.
13. F. Alharbi and A. Vakanski, “Machine learning methods for cancer classification using gene expression data: A review,” *Bioengineering*, vol. 10, no. 2, pp. 1–26, 2023.
14. F. Azuaje, “Artificial intelligence for precision oncology: Beyond patient stratification,” *NPJ Precision Oncology*, vol. 3, no. 1, pp. 1–5, 2019.
15. F. Rohart, B. Gautier, A. Singh, and K. A. Lê Cao, “mixOmics: An R package for ‘omics feature selection and multiple data integration,” *PLOS Computational Biology*, vol. 13, no. 11, pp. 1–19, 2017.
16. F. Sartori, F. Codicè, I. Caranzano, C. Rollo, G. Birolo, P. Fariselli, and C. Pancotti, “A comprehensive review of deep learning applications with multi-omics data in cancer research,” *Genes*, vol. 16, no. 6, pp. 1–34, 2025.
17. G. Kallah-Dagadu, M. Mohammed, J. B. Nasejje, N. N. Mchunu, H. S. Twabi, J. M. Batidzirai, and I. Maposa, “Breast cancer prediction based on gene expression data using interpretable machine learning techniques,” *Scientific Reports*, vol. 15, no. 1, p. 7594, 2025.
18. G. N. Nyakundi, J. Ndiritu, J. M. Ivivi, and T. Kamanu, “Class prediction of high-dimensional data with class imbalance: Breast cancer gene expression data,” *International Journal of Advances in Scientific Research and Engineering*, vol. 10, no. 11, pp. 28–46, 2024.
19. G. Nicora, F. Vitali, A. Dagliati, N. Geifman, and R. Bellazzi, “Integrated multi-omics analyses in oncology: A review of machine learning methods and tools,” *Frontiers in Oncology*, vol. 10, no. 6, pp. 1–11, 2020.
20. H. Chai, X. Zhou, Z. Zhang, J. Rao, H. Zhao, and Y. Yang, “Integrating multi-omics data through deep learning for accurate cancer prognosis prediction,” *Computers in Biology and Medicine*, vol. 134, no. 3, p. 104481, 2021.
21. H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.
22. I. H. Sarker, “Machine learning: Algorithms, real-world applications and research directions,” *SN Computer Science*, vol. 2, no. 3, pp. 1–21, 2021.
23. I. T. Jolliffe and J. Cadima, “Principal component analysis: A review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.
24. J. E. Flores, D. M. Claborne, Z. D. Weller, B. J. M. Webb-Robertson, K. M. Waters, and L. M. Bramer, “Missing data in multi-omics integration: Recent advances through artificial intelligence,” *Frontiers in Artificial Intelligence*, vol. 6, no. 2, pp. 1–15, 2023.
25. J. Labory, E. N. Fotso, and S. Bottini, “Benchmarking feature selection and feature extraction methods to improve the performances of machine-learning algorithms for patient classification using metabolomics biomedical data,” *Computational and Structural Biotechnology Journal*, vol. 23, no. 12, pp. 1274–1287, 2024.
26. J. S. Wekesa and M. Kimwele, “A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment,” *Frontiers in Genetics*, vol. 14, no. 7, pp. 1–11, 2023.
27. K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, “Deep learning–based multi-omics integration robustly predicts survival in liver cancer,” *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018.
28. M. Ali and T. Aittokallio, “Machine learning and feature selection for drug response prediction in precision oncology applications,” *Biophysical Reviews*, vol. 11, no. 1, pp. 31–39, 2018.
29. M. Baptiste, S. S. Moinuddeen, C. L. Soliz, H. Ehsan, and G. Kaneko, “Making sense of genetic information: The promising evolution of clinical stratification and precision oncology using machine learning,” *Genes*, vol. 12, no. 5, pp. 1-15, 2021.
30. M. F. Kabir, T. Chen, and S. A. Ludwig, “A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction,” *Healthcare Analytics*, vol. 3, no. 3–4, pp. 1–9, 2023.
31. M. Greenacre, P. J. F. Groenen, T. Hastie, A. I. D’Enza, A. Markos, and E. Tuzhilina, “Principal component analysis,” *Nature Reviews Methods Primers*, vol. 2, no. 1, pp. 1–24, 2022.
32. M. Sinkala, N. Mulder, and D. Martin, “Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics,” *Scientific Reports*, vol. 10, no. 1, pp. 1–14, 2020.
33. N. Mahendran, P. M. Durai Raj Vincent, K. Srinivasan, and C. Y. Chang, “Machine learning based computational gene selection models: A survey, performance evaluation, open issues, and future research directions,” *Frontiers in Genetics*, vol. 11, no. 12, pp. 1–25, 2020.
34. N. Wang, Q. Zhou, J. Gao, and Z. Wang, “Evaluating the efficacy of PCA and t-SNE in optimizing input features for groundwater level simulation using machine learning models,” *Environmental Earth Sciences*, vol. 84, no. 12, p. 336, 2025.

35. P. K. Kanti, P. Sharma, V. V. Wanatasanappan, and N. M. Said, "Explainable machine learning techniques for hybrid nanofluids transport characteristics: An evaluation of Shapley additive and local interpretable model-agnostic explanations," *Journal of Thermal Analysis and Calorimetry*, vol. 149, no. 21, pp. 1–20, 2024.
36. R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle, "Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets," *Molecular Systems Biology*, vol. 14, no. 6, pp. 1–13, 2018.
37. R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.
38. R. Wang and Z. Wang, "Precision medicine: Disease subtyping and tailored treatment," *Cancers*, vol. 15, no. 15, pp. 1–37, 2023.
39. S. Meshoul, A. Batouche, H. Shaiba, and S. AlBinali, "Explainable multi-class classification based on integrative feature selection for breast cancer subtyping," *Mathematics*, vol. 10, no. 22, pp. 1–27, 2022.
40. S. Sakri and S. Basheer, "Fusion model for classification performance optimization in a highly imbalanced breast cancer dataset," *Electronics*, vol. 12, no. 5, pp. 1–26, 2023.
41. S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," *BMC Bioinformatics*, vol. 7, no. 1, pp. 1–8, 2006.
42. Scikit-learn, "Nested versus non-nested cross-validation," *Scikit-learn*, 2025. Available: https://scikit-learn.org/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html [Accessed by 22/10/2025].
43. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, pp. 1–21, 2015.
44. T. Wang, W. Shao, Z. Huang, H. Tang, J. Zhang, Z. Ding, and K. Huang, "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification," *Nature Communications*, vol. 12, no. 1, pp. 1–13, 2021.
45. X. Lu, X. Lu, Z. Wang, J. D. Iglehart, X. Zhang, and A. L. Richardson, "Predicting features of breast cancer with gene expression patterns," *Breast Cancer Research and Treatment*, vol. 108, no. 2, pp. 191–201, 2007.
46. Y. Dong, S. Zhou, L. Xing, Y. Chen, Z. Ren, Y. Dong, and X. Zhang, "Deep learning methods may not outperform other machine learning methods on analyzing genomic studies," *Frontiers in Genetics*, vol. 13, no. 9, pp. 1–10, 2022.
47. Y. Dou and G. Mirzaei, "MO-GCAN: Multi-omics integration based on graph convolutional and attention networks," *Bioinformatics*, vol. 41, no. 8, pp. 1–9, 2025.
48. Y. Yang and G. Mirzaei, "Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification," *PLOS ONE*, vol. 19, no. 2, pp. 1–17, 2024.
49. Z. Cai, R. C. Poulos, J. Liu, and Q. Zhong, "Machine learning for multi-omics data integration in cancer," *iScience*, vol. 25, no. 2, pp. 1–18, 2022.
50. Z. Momeni, E. Hassanzadeh, M. Saniee Abadeh, and R. Bellazzi, "A survey on single and multi-omics data mining methods in cancer data classification," *Journal of Biomedical Informatics*, vol. 107, no. 6, pp. 1–17, 2020.
51. N. Janakarajan and D. Davidwissel, "SurvBoard," *Kaggle Dataset*, 2024. Available: https://www.kaggle.com/datasets/jnikita/survboard [Accessed by 12/09/2024].