MJoSHT
Malaysian Journal of Science, Health & Technology
[mjosht.usim.edu.my]

USIM

Article

Cancer Classification Challenges in High-Dimensional Microarray Data: An In-Depth Exploration of Machine Learning Models

Wafa' Qasim Al-Jamal[1], Sakinah Ali Pitchay[1,2], Farida Ridzuan[1,2] and Muhammad Harith Noor Azam[1,2]

[1]*Faculty of Science and Technology, Universiti Sains Islam Malaysia. 71800 Nilai, Negeri Sembilan, Malaysia.*

[2] *Cybersecurity & Systems Research Unit, Universiti Sains Islam Malaysia, 71800, Negeri Sembilan, Malaysia.*

Correspondence should be addressed to:

Sakinah Ali Pitchay; sakinah.ali@usim.edu.my

Article Info
Article history:
Received:25 August 2025
Accepted: 15 January 2026
Published:15 Mac 2026

Academic Editor:
Fauziah Abdul Wahid

Malaysian Journal of Science,
Health & Technology

MJoSHT2025, Volume 11, Special Issue on the 5th International Conference on Recent Advancements in Science and Technology (ICoRAST 2025): Responsible Artificial Intelligence – Advancing Science and Technology for Humanity
eISSN: 2601-0003

https://doi.org/10.33102/mjosht.531

***Abstract*— Microarray gene expression profiling has transformed biomedical research by enabling large-scale, parallel analysis of thousands of genes. Despite its promise, cancer classification using Machine Learning (ML) on microarray data continues to face critical challenges, particularly due to high dimensionality, limited sample sizes, and severe class imbalance. These factors contribute to overfitting, poor generalization, and inflated performance metrics, hindering the clinical translation of models. This Structured Literature Review (SLR) examines ML-based cancer classification studies published between 2015 and 2025. This period was marked by the emergence of deep learning, synthetic data generation, and biologically informed modeling. Using a transparent selection protocol, we synthesize findings from over 20 peer-reviewed studies. The review focuses on three methodological pillars: biologically grounded feature selection, constrained data augmentation, and robust performance evaluation. We identify a growing trend toward hybrid feature selection methods that balance statistical relevance and biological interpretability. However, comparative benchmarking across datasets remains limited. Data augmentation techniques, such as Synthetic Minority Oversampling Technique (SMOTE) and Generative Adversarial Networks (GAN)s, are increasingly being adopted. However, they often lack biological validation. This raises concerns about the plausibility of synthetic gene profiles. To address this, we recommend integrating pathway-level constraints and gene ontology checks during the augmentation process. Furthermore, we observe that many studies disproportionately emphasize accuracy. This can misrepresent the model's efficacy in imbalanced settings. Metrics such as Matthews Correlation Coefficient (MCC), F1-score, and precision-recall curves offer more reliable insights. These metrics should be standardized across evaluations. External validation using independent datasets is also essential to assess generalizability. In addition, it helps mitigate dataset-specific bias. Based on the findings, we present a conceptual hybrid framework that integrates biologically informed feature selection, biologically constrained data augmentation, and balanced evaluation protocols. This framework is intended to enhance reproducibility, biological fidelity, and translational reliability in machine learning-based cancer diagnostics, thereby contributing to the advancement of precision oncology.**

***Keywords*— Microarray; high dimensionality; biomedical data; cancer classification**

I. INTRODUCTION

High-throughput genomic technologies have significantly advanced biomedical research, with microarray gene expression profiling emerging as a pivotal tool in oncology. This approach enables the simultaneous analysis of thousands of genes from a single biological sample, offering valuable insights into cancer biology, subtype discovery, drug response prediction, and biomarker identification [1,2]. Despite this potential, microarray data present substantial analytical challenges that hinder their clinical utility, most notably high dimensionality, limited sample sizes, and class imbalance. These issues complicate downstream analysis, often resulting in overfitting, reduced generalization, and unreliable performance in predictive modeling [3, 4]. The imbalance between the number of features and the small number of samples can skew learning algorithms and affect classifier reliability, especially when using traditional machine learning models without appropriate preprocessing steps [5]. Machine Learning (ML) methods such as Support Vector Machines (SVM), Random Forests, and k-Nearest Neighbors have been widely applied to microarray classification. However, their success depends heavily on robust feature selection techniques to reduce dimensionality and enhance interpretability [2, 3]. While Deep Learning (DL) approaches, such as Convolutional Neural Networks (CNNs), have shown promise, they are often limited by small datasets and the risk of generating biologically implausible synthetic samples, particularly when using augmentation techniques such as SMOTE or GANs [6,7]. In this review, the term 'hybrid models' refers to approaches that combine different strategies, such as integrating filter and wrapper feature selection with machine learning classifiers, or combining statistical techniques with biologically informed methods. The phrase constrained (domain-aware) data augmentation refers to augmentation strategies that apply biological or structural constraints (e.g., preserving gene-gene correlations) to ensure that synthetic samples remain realistic and valid. These definitions are used consistently throughout this paper to distinguish between ML, DL, hybrid, and constrained approaches. Moreover, the misuse of evaluation metrics remains a critical issue. Accuracy, though frequently reported, is misleading in imbalanced settings. Metrics such as the F1-score and Matthews Correlation Coefficient (MCC), which better account for the performance of the minority class, remain underutilized in microarray studies. These issues collectively call for a refined analytical framework that integrates both biological relevance and statistical rigour [8].

This review compiles key trends and insights from contemporary research (2015-2025) in ML-based microarray cancer classification. Specifically, it focuses on three innovations: (1) the integration of filter, wrapper, and hybrid feature selection methods with traditional ML classifiers; (2) the adoption of constrained data augmentation techniques that preserve biological structure; and (3) the use of balanced evaluation metrics that reflect model performance under class imbalance. By critically analyzing existing approaches and highlighting methodological gaps, this paper contributes to the development of more accurate, interpretable, and clinically applicable ML models for cancer subtype classification. The insights gained from this review are intended to support future advances in bioinformatics-driven cancer diagnostics and genomic tool development.

Although several review papers have discussed machine learning methods for cancer classification using gene expression data, many of them do not fully capture the methodological depth required for clinical relevance. For example, Ahmad et al. [9] provided a comprehensive overview of ML classification techniques. However, the study did not examine the limitations of relying solely on accuracy metrics or the challenges of biologically unconstrained data augmentation. Similarly, Bhandari et al. [10] focused more on computational learning algorithms than on integrating biological context into feature selection or model evaluation. Meanwhile, Mazlan et al. [11] reviewed various gene selection methods but did not connect these techniques to broader concerns, such as class imbalance or generalizability of performance. In contrast, this paper fills a notable gap by reviewing literature from 2015 to 2025, with a specific emphasis on three underrepresented yet critical aspects: biologically informed feature selection, constrained data augmentation, and robust evaluation metrics.

II RESEARCH GAP

Previous reviews (e.g., [9-11]) have explored machine learning applications in bioinformatics or cancer-related data. Many lack a focused critique of algorithmic challenges specific to high-dimensional datasets or do not incorporate post-2020 methodological advancements. This review narrows that gap by:

(1) targeting studies published between 2015 and 2025 that apply ML to microarray gene expression data with extreme dimensionality,

(2) categorizing findings based on three recurring methodological bottlenecks: feature selection in high-dimensional spaces, imbalance-aware data augmentation, and the use of reliable evaluation metrics, and

(3) offering a comparative synthesis that highlights algorithmic limitations, reproducibility concerns, and areas for improvement in ML model design.

This review contributes to the ML literature by outlining a structured roadmap for developing scalable, interpretable, and generalizable models for complex biological data, particularly when traditional ML pipelines struggle with data sparsity and imbalance.

Microarray gene expression data pose unique challenges for machine learning, primarily due to very high feature dimensionality and small sample sizes. This imbalance can lead to overfitting, poor model generalization, and inflated metrics when classification performance is assessed without proper safeguards. Previous review articles [12, 13] have focused broadly on pipelines or classical issues, such as feature selection. Nevertheless, none have offered a targeted, post-2020 synthesis that explicitly integrates domain-informed feature selection, plausible data augmentation, and balanced evaluation metrics, the three core pillars of this review.

Our work fills this gap by:

- Restricting to the most recent decade (2015-2025),

- Critiquing methodological gaps around augmentation realism and evaluation bias,
- Proposing a systematic hybrid framework for future microarray classification research.

The rest of the paper is organized as follows: Section II reviews recent literature on ML-based microarray cancer classification. Section III explains the methodology and inclusion criteria used to select the studies. Section IV presents a synthesis of the reviewed approaches, comparing how existing methods address challenges in feature selection, augmentation, and evaluation. Sections V and VI present a critical discussion of reviewed approaches, limitations, and concerns regarding interpretability. Section VII suggests a potential research direction. Section VIII concludes the paper with a summary of the findings, a proposed hybrid framework, and recommendations for future research directions.

III. LITERATURE REVIEW

The classification of high-dimensional microarray data has become a significant challenge in data-driven classification tasks, given the large number of gene expression features relative to the small number of available samples. This imbalance introduces several modelling challenges, including overfitting, poor generalizability, and inflated performance estimates from the use of poorly chosen metrics. These challenges are particularly acute in microarray data applications where feature dimensionality may reach tens of thousands, while sample sizes remain in the tens or low hundreds, leading to what is often referred to as the "curse of dimensionality."

Traditional ML classifiers such as Support Vector Machines (SVM), Random Forests (RF), and Extreme Learning Machines (ELM) have demonstrated strong baseline performance when paired with effective preprocessing. Despite this, their effectiveness largely hinges on how well irrelevant or redundant features are eliminated through feature selection. Methods such as the Least Absolute Shrinkage and Selection Operator (LASSO), Minimum Redundancy Maximum Relevance (mRMR), and Principal Component Analysis (PCA) are often employed for feature selection/dimensionality reduction. However, they tend to select features based solely on statistical criteria, potentially overlooking weak yet informative patterns. Mazrua et al. [2] and Deng et al. [3] provided comprehensive surveys of these methods, highlighting their limitations in overseeing high-dimensional biological datasets.

To address this, recent research advocates for biologically inspired or knowledge-integrated feature selection frameworks. For example, hybrid methods that combine filter and wrapper strategies facilitate the selection of features that are both statistically significant and contextually interpretable. Studies such as [14], which used text-mining techniques to prioritise genes based on PubMed literature, offer a model for integrating domain-specific knowledge with algorithmic strategies. These techniques enhance predictive performance and provide transparency and traceability for feature significance, a crucial aspect in critical domains such as biomedical informatics and regulatory ML applications.

In the context of data scarcity, constrained augmentation methods are replacing conventional oversampling techniques. The SMOTE and its various variants are commonly used to rebalance class distributions. However, these methods often rely on simplistic linear interpolations in high-dimensional feature spaces, which can potentially introduce synthetic instances that do not align with the true structure of the data manifold. Studies such as those by Blagus and Lusa [15] and Ke et al. [16] demonstrated that unregulated synthetic sample generation can reduce model generalizability and introduce noise into the learning process.

In response, more recent ML studies have proposed constrained augmentation frameworks in which the generation of synthetic samples is regulated by prior knowledge of gene-gene interactions or by enforcing topological constraints derived from graph-structured embeddings. These strategies aim to generate more plausible synthetic samples that lie within the data manifold while preserving inter-feature dependencies, improving model robustness across test datasets. Ravindran et al. [6] proposed a Wasserstein Tabular GAN (WT-GAN) that incorporates such constraints into its generator loss function to preserve structural dependencies during oversampling. Another critical dimension is performance evaluation. Accuracy has long been used as the go-to metric for classification tasks. However, in class-imbalanced settings, such as microarray data, this metric is notoriously misleading. A classifier can achieve high accuracy by simply favoring the majority class while completely ignoring the minority class, which often represents the target outcome in medical applications.

To address this, robust metrics such as the F1-score, Matthews Correlation Coefficient (MCC), Balanced Accuracy, and Precision-Recall AUC are now being advocated. Chicco and Jurman [17] convincingly argued that MCC is superior in high-imbalance scenarios because it combines all components of the confusion matrix (TP, FP, TN, FN) into a single coefficient. MCC remains meaningful even when class distributions are skewed, unlike accuracy or even raw recall. The F1-score, being the harmonic mean of precision and recall, also plays a central role in balancing the rates of false positives and false negatives. These metrics help avoid misleading conclusions about model performance and are particularly useful in cross-dataset comparisons.

In multi-class classification, which is common in tissue classification or subtype discovery, the macro-F1 and weighted-F1 scores are recommended for assessing class-specific performance [19]. The shift toward these metrics represents a maturation in methodological rigour and reflects a growing understanding of the limitations of oversimplified evaluation strategies.

Lastly, TABLE I in the paper presents a detailed taxonomy of ML methods applied to microarray classification from 2015 to 2025. It synthesizes findings from over 20 key studies, capturing trends in model types, feature selection strategies, augmentation techniques, and performance metrics.

This body of literature reveals a gradual evolution away from purely statistical pipelines toward hybrid, interpretable, and robust machine learning architectures that prioritize generalization and realistic data modelling [20- 22]. However, the fragmentation of approaches and the lack of standardized pipelines for evaluation remain a challenge. Future research is likely to focus on integrating these components into cohesive and reusable machine learning workflows for high-dimensional classification tasks, extending beyond biomedical domains.

TABLE I. TAXONOMY OF ML-BASED METHODS FOR MICROARRAY CANCER CLASSIFICATION (2015–2025)

Ref	Dataset(s)	Data Type	Cancer Type(s)	Feature Selection	Classifier(s)	Data Augmentation	Accuracy	Sample Size	Dimensionality	Pros	Cons
[19]	GEO (Colon, DLBCL, Leukemia, Prostate)	Microarray	Colon, DLBCL, AML/ALL, Prostate	mRMR + MBFA	Ensemble + Attention (SVM, KNN)	No	Up to 100%	62–136	2,000–12,600	Improved accuracy; Efficient	May have scalability issues
[20]	BRCA, BLCA, CIT-Curie	Microarray	Breast, Bladder	CNN + BiGRU	DCGN	SMOTE	94.5% – 99.4%	761–1185	20,000+	Handles subtype diversity	Lower generalizability in sparse data
[4]	Breast, Colon, Leukemia, etc.	Microarray	Multiple types	Isomap + GA	SVM, DNN	No	85.7% –100%	60–203	668–24,188	Robust with few genes	Expensive and dataset-limited
[7]	Arcene, Colon, AML, etc.	Microarray	10+ cancer types	Laplacian Score	CNN	No	Up to 100%	40–200	2,000–12,625	High CNN accuracy	Overfitting in small data
[6]	GEO (Colon, Leukemia, Prostate)	Microarray	Colon, Leukemia, Prostate	Correlation-based	FNN	WT-GAN	~97%	62–102	2,000–6,033	Augmentation improves results	Poor generalization on sparse sets
[5]	GSE45827, GSE14520, etc.	Microarray + RNA-seq	Breast, Liver, etc.	Fuzzy Gene Selection	MLP	No	93%–99%	100–2,000+	Up to 45,782	Early detection enhanced	Normalization inconsistency
[3]	14 GEO datasets	Microarray	Colon, Leukemia, Lung	XGBoost + MOGA	RF, KNN	No	Up to 97%	62–181	>2,000	Removes irrelevant genes	High computational cost
[1]	GSE14520, GSE19804, etc.	Microarray	Breast, CNS, Colon, etc.	PSO	CNN + ASTM	No	95.45%–100%	49–357	2,000–54,676	High accuracy + reduced training	Transfer learning dependency
[21]	TCGA	RNA-seq	33 cancer types	None (CNN-based)	1D/2D CNN	No	95.5% – 95.7%	10,340 tumors	7,100	High precision: marker genes found	Low robustness on rare subtypes
[22]	TCGA, CPTAC-3, DKFZ	RNA-seq	18 cancers	Contrastive Learning	CL-XGBoost	No	AUC 0.8–0.9	308–11,069	20,531	Enhanced representations	Overfitting on small subsets
[23]	GEO (e.g., Breast, Colon)	Microarray	Breast, Colon	ReliefF + LASSO (Hybrid)	SVM, RF	No	91%–97%	70–150	3,000–10,000	Hybrid FS improves interpretability	Longer training time

IV METHODOLOGY

This review adopts a Structured Literature Review (SLR) approach to analyze machine learning-based methods for cancer classification using microarray gene expression data. The goal is to synthesize, evaluate, and compare recent computational methodologies applied to this domain. Notably, studies focusing only on theoretical concepts or addressing non-microarray omics data, such as proteomics, were excluded.

Figure 1 illustrates the flowchart of the reviewing process, while Table 1 summarizes the structured literature review conducted for this study. Literature was sourced from four major databases: PubMed, IEEE Xplore, SpringerLink, and ScienceDirect. The search was guided by relevant keywords, including "microarray," "cancer classification," "machine learning," "deep learning," "feature selection," "gene expression," and "data augmentation." The inclusion criteria were as follows: (1) studies published between 2015 and 2025, (2) application of machine learning or deep learning methods to microarray-based cancer classification, (3) use of feature selection or dimensionality reduction techniques, and (4) reporting of performance metrics beyond simple accuracy. Studies were excluded if they lacked experimental validation or if their scope was limited to theoretical or conceptual analysis.

In total, more than 20 peer-reviewed journal articles and conference papers were selected for review. The selected studies were then analyzed and grouped based on three recurring methodological themes: (1) feature selection strategies; (2) data augmentation methods; and (3) performance evaluation metrics. This taxonomy served as the foundation for a comparative framework used to identify common limitations, cross-study observations, and future directions.

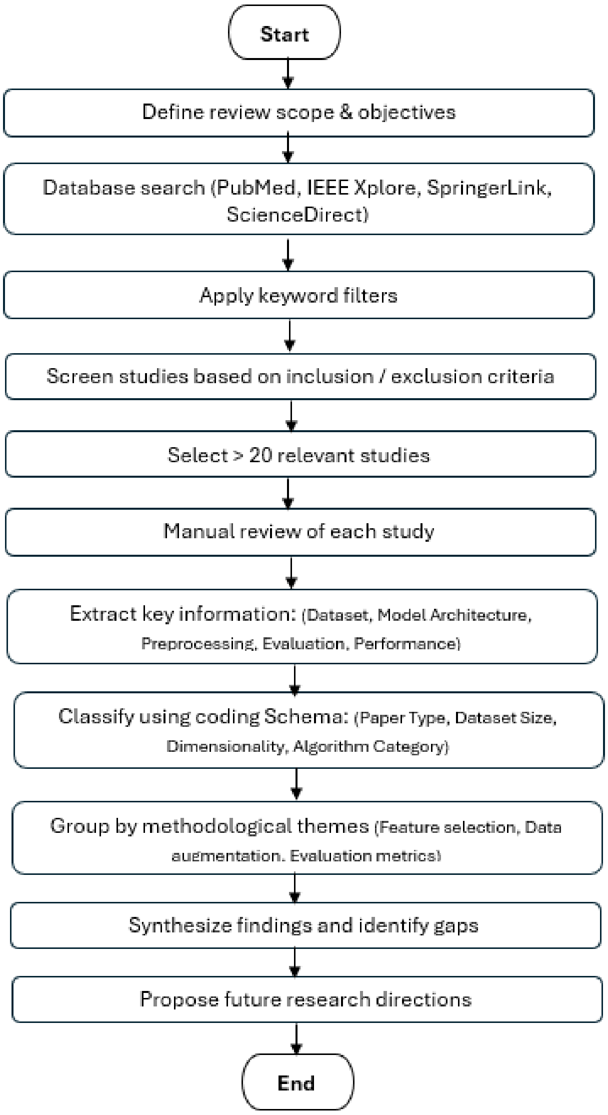
Start
Define review scope & objectives
Database search (PubMed, IEEE Xplore, SpringerLink, ScienceDirect)
Apply keyword filters
Screen studies based on inclusion / exclusion criteria
Select > 20 relevant studies
Manual review of each study
Extract key information: (Dataset, Model Architecture, Preprocessing, Evaluation, Performance)
Classify using coding Schema: (Paper Type, Dataset Size, Dimensionality, Algorithm Category)
Group by methodological themes (Feature selection, Data augmentation, Evaluation metrics)
Synthesize findings and identify gaps
Propose future research directions
End

Figure 1. Flow chart of structured literature review

Each selected study was manually reviewed to extract information about the dataset(s) used, model architecture, preprocessing pipeline, evaluation strategy, and reported performance. To ensure consistency, a coding schema was used to classify papers by type (e.g., hybrid feature selection, constrained augmentation), dataset size, dimensionality, and algorithm category (e.g., SVM, CNN, ensemble).

Another critical issue is the widespread misuse of accuracy as a sole performance metric in imbalanced datasets. In many cancer-related datasets, the minority class, often the disease of interest, holds greater clinical importance. However, high accuracy can mask the model's failure to detect rare yet vital patterns. [17, 24] have suggested that relying solely on accuracy results in inflated and misleading evaluations. More informative metrics, such as the F1-score and Matthews Correlation Coefficient (MCC), are required to assess classifier performance more effectively in these settings.

Feature selection is another area of concern. Traditional filter-based methods, such as PCA and mRMR, prioritize statistical relevance, often discarding weakly expressed yet biologically important genes. It is essential to exercise caution, as such actions may compromise the model's interpretability and degrade its performance [25]. In contrast, wrapper-based or hybrid methods, especially those integrating biological domain knowledge, have demonstrated greater success in preserving interpretability and enhancing prediction accuracy [26].

TABLE II. STRUCTURED LITERATURE REVIEW SUMMARY

Component	Details
Review Type	Structured Literature Review (SLR)
Objective	Synthesize, evaluate, and compare ML methods for cancer classification using microarray data.
Databases searched	PubMed, IEEE Xplore, SpringerLink, ScienceDirect
Search keywords	"microarray," "cancer classification," "machine learning," "deep learning," "feature selection," "gene expression," "data augmentation"
Inclusion criteria	1. Published between 2015 and 2025. 2. ML/DL applied to microarray-based cancer classification. 3. Use of feature selection/dimensionality reduction. 4. Performance metrics beyond accuracy.
Exclusion criteria	• No experimental validation. • Purely theoretical. • Non-microarray omics (e.g., proteomics).
Number of studies	> 20 peer-reviewed journal articles and conference papers.
Analytical Themes	• Feature selection strategies. • Data augmentation method. • Performance evaluation metrics.
Classification Schema	• Paper type (e.g., hybrid feature selection). Dataset size. • Dimensionality. • Algorithm category (e.g., SVM, CNN, ensemble).

Collectively, these challenges underscore the importance of biologically informed and statistically rigorous modeling practices in microarray-based cancer classification. Addressing overfitting, adopting robust, balanced evaluation metrics, using biologically guided feature selection, and applying realistic augmentation techniques are all essential steps toward building reliable, clinically applicable ML models.

To manage class imbalance, many studies have employed synthetic data generation methods, such as SMOTE and GANs. While these techniques can improve minority class representation, they often produce unrealistic gene expression patterns if applied without constraints [15]. Bagui and Li [27] emphasized that excessive reliance on synthetic data can reduce generalizability and introduce noise. This highlights the need for constrained augmentation strategies that balance class distributions while preserving biological plausibility.

Instead of merely summarizing findings, the review organizes and critiques studies through the lens of three core challenges:

- *Feature Selection in High-Dimensional Spaces*: We evaluated whether methods relied purely on statistical filters or integrated domain knowledge, such as graph-based or literature-based filtering (e.g., PubMed prioritization).
- *Data Augmentation Strategies*: We assessed whether augmentation methods preserved structural relationships between features or simply interpolated in feature space without constraints.
- *Evaluation Metrics*: We tracked the prevalence of balanced metrics, such as MCC, F1-score, and PR-AUC across datasets, and whether authors benchmarked models beyond accuracy.

Despite notable progress in ML-based cancer classification, several methodological limitations persist, affecting model reliability and clinical relevance. A significant concern is overfitting, especially in microarray experiments where the number of genes (features) far exceeds the number of samples. As noted by Kim and Pang [28], overfitting occurs when models capture noise instead of true biological signals, leading to inflated performance during training yet poor generalization in real-world scenarios. This problem is particularly acute in deep learning architectures trained with limited data and weak regularization [29].

Despite methodological progress, certain patterns of weakness were consistently observed:

- Overfitting remains prevalent due to the extremely small sample sizes in many studies, especially those employing deep learning models.
- Feature selection techniques that rely on variance thresholds or mutual information often discard subtle but informative patterns.
- Many augmentation methods generate synthetic instances that do not reflect valid feature interdependencies.
- A significant proportion of studies report accuracy as the primary performance metric, which is misleading in class-imbalanced datasets.

By systematically categorizing, comparing, and synthesizing ML methods applied to microarray classification tasks, this review offers a reproducible and extendable methodology for future computational meta-analyses in high-dimensional domains. The selected methodology identifies methodological gaps and provides a foundation for evaluating proposed hybrid frameworks in subsequent sections. This section is designed to align with the extended literature review and discussion sections, ensuring consistency and traceability between methodological setup and resulting synthesis.

V Synthesis of Reviewed Approaches

Drawing on the selected literature, this section examines how recent studies align with core ML challenges in microarray classification. Rather than presenting new experimental results, we critically examine how existing methodologies align with the three core pillars of an emerging hybrid framework: biologically informed feature selection, constrained data augmentation, and robust performance evaluation. First, a significant proportion of the reviewed studies employ statistical feature selection methods, such as mRMR, LASSO, or PCA. However, only a minority integrates biological knowledge during the gene filtering process. Wu et al. [14] introduced a text-mining approach to prioritize biologically relevant genes using PubMed evidence. Similarly, hybrid strategies that combine statistical and domain knowledge, as demonstrated by Chiew et al. [23] and Qiao et al. [30], offered improved accuracy without sacrificing interpretability. These approaches represent a meaningful shift toward biologically grounded gene selection.

Second, while oversampling techniques such as SMOTE and GANs are commonly employed to address class imbalance, many studies apply them without constraints, thereby risking the generation of biologically implausible gene profiles. Studies such as Ke et al. [16] emphasized the significance of constrained (domain-aware) augmentation strategies that enforce biological plausibility in synthetic instances, thereby improving generalizability and preserving data fidelity.

Third, evaluation metrics remain a persistent concern. Despite its popularity, accuracy alone is often insufficient in imbalanced settings. More robust metrics, such as the F1-score and Matthews Correlation Coefficient (MCC), recommended by Chicco and Jurman [17] and Alsaeedi [31], offered better insights into model performance, especially for underrepresented classes. However, their use remains limited across the reviewed literature.

While progress has been made in each of these areas, only a few studies have successfully integrated all three components into a cohesive framework. This synthesis underscores the need for structured hybrid ML approaches that improve predictive accuracy, ensure biological plausibility, clinical relevance, and reliable evaluation across diverse microarray datasets.

We present a conceptual hybrid framework that combines best practices in preprocessing, data augmentation, and evaluation. The framework is illustrated in Figure 2, which highlights the integration of biologically informed feature selection, constrained augmentation, and robust evaluation into a scalable pipeline.

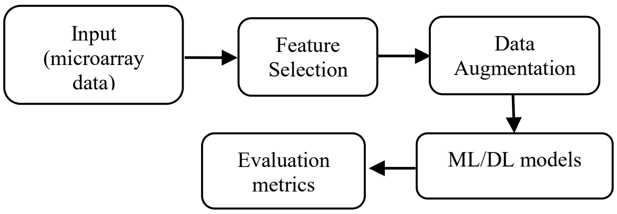

Figure 2. Conceptual hybrid framework for microarray-based cancer classification.

VI DISCUSSION

The classification of high-dimensional microarray data remains a significant challenge in machine learning research, particularly when balancing model accuracy, interpretability, and generalizability across diverse datasets. This review consolidates advances in ML frameworks that pivot away from purely deep learning-centric approaches, instead emphasizing hybrid pipelines that address key issues such as class imbalance, overfitting, and inconsistent evaluation. A recurring issue is the discrepancy between reported accuracy and actual generalizability. Studies such as Chicco and Jurman [17] and He et al. [32] demonstrate that, despite achieving high training accuracy, many models fail on unseen datasets due to overfitting in small, high-dimensional gene spaces. This problem is exacerbated by augmentation methods such as SMOTE and GANs, which, if unconstrained, produce synthetic data that lacks statistical or structural coherence [6, 15].

Several recent studies emphasize the importance of domain-aware or constrained data augmentation frameworks [16, 28], arguing for augmentation techniques that preserve class boundaries and minimise noise injection. Without such constraints, models risk learning spurious correlations or noise patterns that fail to generalize. In feature selection, hybrid techniques that combine statistical and heuristic methods have shown promise. For example, [3] integrates XGBoost and genetic algorithms to enhance gene selection, achieving improved performance across colon and leukaemia datasets. These approaches reduce dimensionality and maintain model interpretability, which is crucial in high-stakes medical domains. Notably, Li et al. [33] and Pani 34] supported the use of hybrid filter-wrapper models that retain low-variance yet meaningful features. This mitigates the risk of discarding weak signals, a common issue with methods such as PCA and mRMR alone.

Evaluation remains another persistent gap. While accuracy remains the default metric, it fails when class distributions are skewed. Balanced metrics, such as F1-score and Matthews Correlation Coefficient (MCC), provide more nuanced insights. As illustrated in Figure 3, most reviewed studies still report accuracy as their primary metric, with far fewer adopting F1-score or MCC. This imbalance highlights the ongoing need for standardised, balanced evaluation protocols, particularly when minority groups carry a higher diagnostic burden [17, 31]. Nevertheless, the adoption of MCC and other balanced metrics remains limited, despite consensus on their effectiveness [17, 26]. Additionally, standardised benchmarking protocols are often lacking, rendering cross-study comparisons unreliable. Frameworks such as Iso-G [4] and ReliefF-LASSO [23] demonstrated the potential of reproducible pipelines. However, few studies have tested models across diverse or external datasets. The lack of external validation severely undermines claims of model robustness.

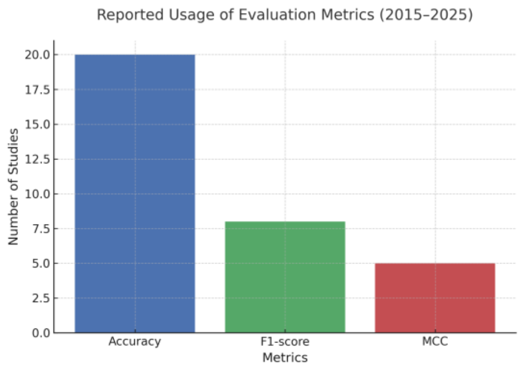

Figure 3. Metric Usage Trends in ML-based microarray cancer classification studies (2015–2025)

Emerging work has also highlighted interpretability. While DL models such as CNNs [7, 20] boast high accuracy, their "black-box" nature makes them challenging to validate and trust, especially without saliency maps or feature attribution techniques. Shifting toward explainable AI (XAI) is a critical next step for clinical acceptability. Thus, this review emphasizes a methodological shift: from accuracy-centric pipelines to transparent, interpretable, and reproducible ML frameworks. Rather than introducing novel algorithms, most performance gains arise from thoughtful integration of hybrid feature selection, constrained augmentation, balanced evaluation metrics, and standardized testing protocols. Thus, future work should develop open-source benchmarking suites for microarray data to enable robust testing across datasets, feature spaces, and class imbalances. Collaborative initiatives between machine learning engineers and domain scientists will be essential for ensuring statistical validity aligns with clinical relevance.

VII POTENTIAL RESEARCH DIRECTIONS

Based on the findings and gaps identified in this review, several future directions can advance the state of ML-based microarray data classification:

1. *Advanced Feature Selection Algorithms for High-Dimensional Spaces*

There is still a need for improved methods to select the most relevant features when working with extremely high-dimensional microarray data. Future work should explore hybrid feature selection methods that combine filters, wrappers, and embedded techniques to improve both performance and interpretability.

More controlled data augmentation techniques include common methods such as SMOTE and GANs. Despite their helpfulness, they often produce unrealistic or noisy samples. Upcoming research should investigate smarter, adaptive techniques that preserve the structure of real data and minimise the introduction of noise, especially in sensitive classification tasks.

2. *Benchmarking Frameworks Across Multiple Datasets*

One current limitation is the lack of consistency in how models are evaluated. Future studies should prioritize developing shared benchmarking frameworks that allow researchers to test their models on multiple datasets using the same metrics and protocols for fairer comparisons.

3. *Integrated ML Pipelines for Microarray Classification*

Rather than building models in a piecemeal manner, future research should focus on end-to-end pipelines that encompass all aspects, including feature selection, data balancing, training, and evaluation. This helps improve reproducibility and ensures all steps are aligned.

4. *Improving the Interpretability of ML Models*

With so many features, it is easy for ML models to become black boxes. Research should focus on building models that perform well and provide understandable explanations, either through explainable architectures or post-hoc tools.

5. *Learning Better from Limited Data*

Microarray datasets often have very few samples. Using techniques such as transfer learning or meta-learning from related tasks can help models generalise better, even with limited labelled data.

VIII CONCLUSIONS AND FUTURE WORK

This review provides an in-depth exploration of machine learning (ML) models for classifying high-dimensional microarray gene expression data in cancer research, a domain characterised by significant challenges, including feature sparsity, limited sample sizes, and severe class imbalance. Drawing on literature published between 2015 and 2025 across PubMed, IEEE Xplore, SpringerLink, and ScienceDirect, we synthesized prevailing trends, methodological gaps, and emerging innovations shaping the current landscape of ML-based cancer classification.

In contrast to earlier reviews that predominantly emphasized classifier performance in isolation, our analysis brings attention to three critical yet underexplored dimensions: (1) biologically informed feature selection strategies, (2) statistically constrained data augmentation techniques, and (3) robust, class-sensitive evaluation metrics. This tripartite focus provides a technically grounded framework that effectively bridges algorithmic sophistication with practical application in biomedical contexts.

A recurring limitation is the overreliance on accuracy as a performance indicator. Many studies lack rigorous cross-validation, external benchmarking, and class-wise performance analysis, leading to inflated results and limited generalizability. Balanced metrics such as the F1-score and Matthews Correlation Coefficient (MCC) offer more reliable insights, underscoring the need for standardized evaluation protocols.

We also highlight the promise of hybrid feature selection approaches that integrate statistical relevance with domain-specific insights. Techniques such as ReliefF combined with LASSO, or PubMed-informed gene filtering, demonstrate potential in producing compact yet biologically meaningful feature sets. However, as datasets grow in scale, automated, transparent selection mechanisms that preserve reproducibility are urgently needed.

To address these gaps, we propose a conceptual hybrid ML framework that integrates best practices in preprocessing, data augmentation, and evaluation, as displayed in Figure 2. This framework is algorithm-agnostic, adaptable to diverse high-dimensional data types, and provides a scalable blueprint for bioinformatics and precision oncology. Looking ahead, several avenues warrant further investigation:

- Cross-dataset evaluation: Future studies should benchmark models across multiple publicly available datasets to assess generalizability and mitigate dataset-specific biases. External validation using unseen data should become a standard in performance reporting.
- Adaptive augmentation pipelines: Research should prioritize augmentation methods that are sensitive to class boundaries, feature correlations, and distributional properties, thereby reducing the risk of generating unrealistic samples.
- Modular and reproducible ML pipelines: There is a clear opportunity to develop unified pipelines that integrate preprocessing, dimensionality reduction, model selection, hyperparameter tuning, and evaluation into reproducible workflows.
- Explainability and transparency: In high-stakes domains such as cancer diagnostics, interpretability is paramount. Future models should incorporate inherently interpretable architectures or post-hoc explanation tools (e.g., SHAP, LIME) to facilitate the elucidation of decision processes.
- Transfer and meta-learning: Given the scarcity of labeled microarray data, leveraging transfer learning from related omics domains and exploring meta-learning strategies may enhance adaptability and performance on small datasets.

Although significant advancements have been made in applying machine learning to cancer classification challenges within microarray data, ongoing methodological limitations continue to hinder its translational efficacy. This review outlines these challenges and proposes a structured roadmap to facilitate the development of more robust, interpretable, and transferable machine learning models. The advancement of this field will rely on interdisciplinary collaboration, rigorous benchmarking processes, and a steadfast commitment to methodological transparency.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

ACKNOWLEDGEMENT

We would like to thank the Faculty of Science and Technology for the facilities and support provided.

REFERENCES

[1] N. Alrefai, O. Ibrahim, H. M. F. Shehzad, A. Altigani, W. Abu-ulbeh, M. Alzaqebah, and M. K. Alsmadi, “An integrated framework based deep learning for cancer classification using microarray datasets,” *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 3, pp. 2249–2260, 2023.

[2] H. AlMazrua and H. Alshamlan, “A comprehensive survey of recent hybrid feature selection methods in cancer microarray gene expression data,” *IEEE Access*, vol. 10, pp. 71427–71449, 2022.

[3] X. Deng, M. Li, S. Deng, and L. Wang, “Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification,” *Med. Biol. Eng. Comput.*, vol. 60, no. 3, pp. 663–681, 2022.

[4] L. Wang, X. Chen, and Y. Li, “Integrative dimensionality reduction using Isomap and Genetic Algorithms (Iso-GA) for robust cancer

classification from microarray data," *BMC Bioinformatics*, vol. 24, no. 1, p. 112, 2023.

[5] M. Khalsan, A. Karimi, and E. Amirzadeh, "Fuzzy gene selection coupled with multilayer perceptrons for early cancer detection using microarray data," *J. Biomed. Inform.*, vol. 138, p. 104287, 2023.

[6] S. Ravindran, K. Ramesh, and P. Balasubramanian, "Wasserstein Tabular GAN (WT-GAN) for addressing class imbalance in microarray-based cancer classification," *Sci. Rep.*, vol. 14, no. 1, p. 5678, 2024.

[7] S. H. Shah, M. J. Iqbal, I. Ahmad, S. Khan, and J. J. Rodrigues, "Optimized gene selection and classification of cancer from microarray gene expression data using deep learning," *Neural Comput. Appl.*, pp. 1–12, 2020.

[8] A. B. I. Issa, "Exploring the transformative impact of AI across industries and its role in shaping global advancements," *Univ. J. Future Impact Artif. Intell.*, vol. 1, no. 1, Art. no. 24, 2024.

[9] F. Alharbi and A. Vakanski, "Machine learning methods for cancer classification using gene expression data: A review," *Bioengineering*, vol. 10, no. 2, p. 173, 2023.

[10] N. Bhandari, R. Walambe, K. Kotecha, and S. P. Khare, "A comprehensive survey on computational learning methods for analysis of gene expression data," *Front. Mol. Biosci.*, vol. 9, p. 907150, 2022.

[11] A. U. Mazlan, N. A. Sahabudin, M. A. Remli, N. S. N. Ismail, M. S. Mohamad, H. W. Nies, and N. B. Abd Warif, "A review on recent progress in machine learning and deep learning methods for cancer classification on gene expression data," *Processes*, vol. 9, no. 8, p. 1466, 2021.

[12] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.

[13] A. M. Musolf, E. R. Holzinger, J. D. Malley, and J. E. Bailey-Wilson, "What makes a good prediction? Feature importance and beginning to open the black box of machine learning in genetics," *Hum. Genet.*, vol. 141, no. 9, pp. 1515–1528, 2022.

[14] G. Wu, A. Zaker, A. Ebrahimi, S. Tripathi, and A. Mer, "Text-mining-based feature selection for anticancer drug response prediction," *Bioinform. Adv.*, vol. 4, no. 1, p. vbae047, 2024.

[15] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," *BMC Bioinformatics*, vol. 14, no. 1, 2013.

[16] P. F. Ke, D. S. Xiong, J. H. Li, Z. L. Pan, J. Zhou, S. J. Li, J. Song, X. Y. Chen, G. X. Li, J. Chen, and X. B. Li, "An integrated machine learning framework for a discriminative analysis of schizophrenia using multi-biological data," *Sci. Rep.*, vol. 11, no. 1, p. 14636, 2021.

[17] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, 2020.

[18] B. J. Parker, S. Günter, and J. Bedo, "Stratification bias in low signal microarray studies," *BMC Bioinformatics*, vol. 8, no. 1, p. 326, 2007.

[19] W. Xie, L. Wang, K. Yu, T. Shi, and W. Li, "Improved multi-layer binary firefly algorithm for optimizing feature selection and classification of microarray data," *Biomed. Signal Process. Control*, vol. 79, p. 104080, 2023.

[20] J. Shen, J. Shi, J. Luo, H. Zhai, X. Liu, Z. Wu, C. Yan, and H. Luo, "Deep learning approach for cancer subtype classification using high-dimensional gene expression data," *BMC Bioinformatics*, vol. 23, no. 1, p. 430, 2022.

[21] M. Mostavi, P. Mirbagheri, and C. Wang, "Convolutional neural network models for cancer type prediction based on gene expression," *BMC Med. Genomics*, vol. 13, no. 1, p. 134, 2019.

[22] A. Sun, E. J. Franzmann, Z. Chen, and X. Cai, "Deep contrastive learning for predicting cancer prognosis using gene expression values," *Brief. Bioinform.*, vol. 25, no. 6, p. bbae544, 2024.

[23] K. Chiew, C. Tan, K. Wong, K. Yong, and W. Tiong, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Inf. Sci.*, vol. 484, pp. 153–166, 2019.

[24] M. Buda, A. Maki, and M. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Netw.*, vol. 106, pp. 249–259, 2018.

[25] Y. Ju, L. Li, L. Jiao, Z. Ren, B. Hou, and S. Yang, "Modified diversity of class probability estimation co-training for hyperspectral image classification," *arXiv preprint arXiv:1809.01436*, 2018.

[26] M. Bai, J. Liu, Z. Long, J. Luo, and D. Yu, "A comparative study on class-imbalanced gas turbine fault diagnosis," *Proc. Inst. Mech. Eng., Part G: J. Aerosp. Eng.*, vol. 237, no. 3, pp. 672–700, 2023.

[27] S. Bagui and K. Li, "Resampling imbalanced data for network intrusion detection datasets," *J. Big Data*, vol. 8, no. 1, p. 6, 2021.

[28] M. Kim and P. Kang, "Text embedding augmentation based on retraining with pseudo-labeled adversarial embedding," *IEEE Access*, vol. 10, pp. 8363–8376, 2022.

[29] P. Yao, S. Shen, M. Xu, P. Liu, F. Zhang, J. Xing, P. Shao, B. Kaffenberger, and R. X. Xu, "Single model deep learning on imbalanced small datasets for skin lesion classification," *IEEE Trans. Med. Imaging*, vol. 41, no. 5, pp. 1242–1254, 2021.

[30] Y. Qiao, Y. Xiong, H. Gao, X. Zhu, and P. Chen, "Protein-protein interface hot spots prediction based on a hybrid feature selection strategy," *BMC Bioinformatics*, vol. 19, no. 1, 2018.

[31] A. H. Alsaeedi, H. H. R. Al-Mahmood, Z. F. Alnaseri, M. R. Aziz, D. Al-Shammary, A. Ibaida, and K. Ahmed, "Fractal feature selection model for enhancing high-dimensional biological problems," *BMC Bioinformatics*, vol. 25, no. 1, p. 12, 2024.

[32] W. He, H. Huang, X. Chen, J. Yu, J. Liu, X. Li, H. Yin, K. Zhang, and L. Peng, "Radiomic analysis of enhanced CMR cine images predicts left ventricular remodeling after TAVR in patients with symptomatic severe aortic stenosis," *Front. Cardiovasc. Med.*, vol. 9, p. 1096422, 2022.

[33] Z. Li, W. Xie, and T. Liu, "Efficient feature selection and classification for microarray data," *PLoS One*, vol. 13, no. 8, e0202167, 2018.

[34] S. Pani, B. Ratha, and A. Mishra, "Performance analysis of microarray data classification using machine learning techniques," *Int. J. Knowl. Discov. Bioinform.*, vol. 5, no. 2, pp. 43–54, 2015.