PENINSULA PUBLISHING LLC

EDRAAK
Vol. **(2026)**, 2026, **pp**. 148–157
ISSN: 3078-8412

Review Article

Machine Learning Techniques for Breast Cancer Prediction in Imbalanced Datasets: A Systematic Review

Omega John Unogwu [1,*], Andrea Ngohide Abaagu [2], Ankar Tersoo Catherine [3]

[1] *National Space Research and Development Agency, Nigeria*

[2] *Department of Medicine and Surgery, Benue State University, Makurdi, Nigeria*

[3] *College of Education, Oju, Benue State, Nigeria*

ARTICLEINFO

Article History
Received 05 Jun 2026
Accepted: 26 Jul 2026
Revised 25 Aug 2026
Published 10 Sep 2026

Keywords
Breast Cancer,
Class Imbalance,
Machine Learning,
Deep Learning,
Classification,
Resampling,
Evaluation Metrics,
Systematic Review

ABSTRACT

Background: Class imbalance remains a central methodological obstacle in machine learning (ML)-based breast cancer prediction, where the clinically critical minority class is systematically underrepresented relative to the majority class. A systematic review by Ghavidel and Pazos synthesized ML approaches to this problem across breast cancer prediction literature published between 2008 and 2023. Given the continued rapid publication of ML and deep-learning methods for breast cancer classification, detection, and prognosis since that search closed, an updated synthesis is warranted. Objective: This manuscript reports the protocol, conceptual framework, and methodology for a systematic review that updates and extends the prior evidence synthesis, focusing on studies published between September 2023 and August 2026 that apply ML techniques to breast cancer prediction under conditions of class imbalance. Methods: The review follows PRISMA 2020 reporting guidance. A multi-database search strategy, eligibility criteria, and a standardized data-extraction framework, harmonized where possible with the prior review, are defined and reported transparently, together with an expanded conceptual background covering the taxonomy of imbalance-handling strategies and the evaluation metrics appropriate to imbalanced classification. Screening and extraction of individual studies, and the resulting comparative synthesis, dataset and algorithm characterization, quality assessment, and discussion, will be completed and reported in subsequent parts of this manuscript once the documented multi-database search has been executed and verified; no screening, inclusion, or performance figures are reported until they can be confirmed against source records. Conclusion (protocol stage): This manuscript establishes a transparent, reproducible, non-fabricated methodological and conceptual foundation for evaluating how the recent literature has addressed class imbalance in breast cancer prediction.

1. INTRODUCTION

Female breast cancer is the most commonly diagnosed cancer among women worldwide and remains the leading cause of cancer death in women, a pattern confirmed by the 2022 GLOBOCAN estimates [1] and reaffirmed in the subsequent 2024 update [2]. These estimates also document substantial variation in incidence and mortality across world regions and levels of human development, reflecting differences in screening infrastructure, risk-factor prevalence, and access to timely diagnosis and treatment [1], [2]. Reducing breast-cancer mortality depends heavily on early detection, accurate diagnostic classification, and reliable risk and prognosis prediction, and machine learning (ML) has become an increasingly common tool for supporting each of these tasks from mammographic screening and histopathological classification to genomic risk stratification and survival prediction.

Machine learning has become an increasingly prominent component of breast cancer care across the full continuum from population screening to survivorship. In imaging-based screening and diagnosis, computer-aided detection and diagnosis (CAD) systems built on classical ML and, increasingly, deep convolutional and transformer architectures, are used to flag suspicious mammographic, ultrasound, or MRI findings for radiologist review. A 2025 systematic review synthesizing 50 studies published between 2018 and 2025 found that optimized classical models, deep-learning models, and hybrid/ensemble architectures have each, in different studies, reported near-ceiling accuracy on individual public

*Corresponding author email: unogwuomega@gmail.com
DOI: https://doi.org/10.70470/EDRAAK/2026/012

mammography benchmarks, while also cautioning that most of this work remains confined to binary classification tasks evaluated on relatively curated, single-institution datasets rather than heterogeneous, real-world screening populations [3]. In histopathology, genomics, and electronic health record data, ML is applied to diagnostic subtyping, molecular risk stratification, and survival or recurrence prognosis. Across all of these applications, however, the datasets available for model development and evaluation are rarely balanced between outcome classes, which motivates the specific methodological focus of this review.

A dataset is described as imbalanced when its outcome categories are not approximately equally represented [4]. The imbalance ratio (IR) for binary classification problems equals the number of majority-class examples divided by the number of minority-class examples although different research papers use various methods to present this data including showing minority-class percentages; this review extracts and reports whichever representation each source study provides (Table III) rather than imposing a single convention retroactively. The imbalance situation extends beyond being a matter which only affects academic studies. In an influential review of imbalanced learning, He and Garcia note that standard learning algorithms are implicitly biased toward whichever class dominates the training distribution, because the optimization objectives typically used — most simply, minimizing overall classification error — treat every misclassification as equally costly regardless of which class it involves [5]. In breast cancer applications this typically, though not universally, takes the form of a numerically dominant “healthy,” “benign,” or “no-recurrence” class and a smaller, clinically critical class corresponding to malignancy, recurrence, or a rare molecular subtype; in some prognostic or subtype-classification tasks a reverse or more complex multi-class imbalance can occur, and this review does not assume a fixed direction of skew for every dataset.

A substantial methodological literature has developed strategies to counteract this majority-class bias, which following established taxonomies [4], [6] this review organizes into three broad families, synthesized in full in Section IV (Part 2). Data-level methods modify the training-set class distribution prior to model fitting, either by oversampling the minority class (through simple duplication, or synthetic generation via the Synthetic Minority Over-sampling Technique, SMOTE, which interpolates new minority-class instances between existing neighbors [7], or its adaptive variant ADASYN, which concentrates synthetic-sample generation near the decision boundary rather than uniformly across the minority class [8]), by undersampling the majority class (through random removal, or informed methods such as Tomek-link removal, which deletes majority-class instances lying immediately adjacent to minority-class instances and thereby sharpens the class boundary [9]), or through a hybrid combination of both. Algorithm-level methods instead modify the learning algorithm itself through class weighting, cost-sensitive loss functions, or decision-threshold adjustment so that errors on the minority class are penalized more heavily without altering the underlying data distribution. Ensemble and hybrid approaches embed data-level or algorithm-level imbalance handling inside a bagging or boosting framework; a widely cited taxonomic review of this family documented substantial empirical benefit over single-classifier baselines across a broad range of application domains, while also cautioning that no single ensemble strategy dominates uniformly across all imbalance conditions [6].

The practical consequence of ignoring imbalance is most easily seen through a hypothetical, purely illustrative example rather than any single reported study. The dataset used for screening contains 95% benign samples together with 5% malignant samples. The classifier which produces a constant output of "benign" for all samples would reach 95% total accuracy but it would fail to detect any malignant cases which results in 0% sensitivity and creates an unacceptable clinical situation despite the high accuracy number. This situation which people sometimes call the "accuracy paradox" demonstrates how evaluation metric selection becomes a biased process when dealing with imbalanced classes which led to the establishment of accuracy-only reporting as a separate methodological weakness in this review (Table III; Section VI).

Multiple additional metrics which work better with unbalanced data sets appear in the review's extraction framework to track their values. The medical community uses Sensitivity (recall) as their main diagnostic screening metric because it shows how well the test detects actual positive cases (e.g., malignant cases) which determines the number of false-negative results. The Specificity metric which equals TN / (TN + FP) shows how well the system detects negative cases and it affects the false-positive rate together with its resulting expenses which include unnecessary biopsies and patient distress and extra medical tests. Precision, TP / (TP + FP), measures the proportion of positive predictions that are correct. The F1-score which represents the harmonic mean of precision and recall values creates a single value to show these metrics but it becomes inaccurate when dealing with highly unbalanced data because it does not consider true negative results. Balanced accuracy, the unweighted average of sensitivity and specificity, eliminates the effects which different class sizes have on regular accuracy measurements. The ROC-AUC function displays how sensitivity numbers affect false positive rate values when using different classification thresholds and PR-AUC shows how precision numbers affect recall values; research indicates that PR-AUC provides better information than ROC-AUC when dealing with scarce positive cases because ROC curves create similar-looking patterns between different classifiers which maintain different precision levels resulting in varying actual false-positive rates [10]. The Matthews correlation coefficient (MCC) calculates values from all four position cells in the confusion matrix to create a single metric which studies have proven to be more reliable than accuracy and F1 when dealing with unbalanced classes [11]. The review documents all metric information from every included study in Table III instead of showing only the primary metric which each study's authors chose to emphasize.

Breast cancer datasets show class imbalance because multiple independent factors create this condition which includes the natural occurrence of cancer in populations used for screening and diagnosis and the original development process of public benchmark datasets and the uncommon occurrence of particular pathological and molecular tumor types and the choice between binary and multi-class prediction tasks. Illustratively, a recent comparative study of resampling techniques evaluated across five cancer datasets including a breast-cancer-specific prognostic dataset with a reported outcome distribution of roughly 84% “alive” versus 16% “deceased” at five years found that hybrid resampling methods combined with Random Forest classifiers produced the strongest mean performance across that benchmark [12]. Imbalance is not confined to tabular or classical ML settings: comparative work applying deep convolutional networks directly to full-field digital mammography found that a greater degree of class imbalance was associated with a stronger bias toward the majority class, that this bias could be partly counteracted by standard imbalance techniques adapted to the deep-learning setting (class weighting, oversampling, undersampling, and synthetic lesion generation), but also that undersampling could measurably reduce ROC-AUC under severe imbalance a reduction of 0.066 AUC was reported at a 19:1 benign-to-malignant ratio in one of the evaluated datasets illustrating that imbalance-handling techniques carry their own performance trade-offs and are not uniformly beneficial [13]. The review does not establish a universal ratio for class imbalance or set severity limits because it retrieves and displays dataset class distributions from studies which offer enough data to determine these distributions.

The review establishes separate categories for breast cancer risk prediction and diagnosis/classification and detection and prognosis prediction tasks. The breast cancer risk prediction task involves predicting disease onset probability for people who do not have cancer yet by using their demographic and hereditary and lifestyle information. The diagnosis/classification process requires medical professionals to differentiate between malignant and benign findings through feature analysis of imaging results and biopsy samples. The detection task requires medical professionals to find suspicious lesions through imaging which serves as the initial stage for further diagnostic classification. The prognosis task requires medical professionals to forecast patient survival and treatment results and cancer recurrence by using clinical data and pathological information and molecular test results which doctors obtain during patient diagnosis. The review presents these tasks as separate entities because they serve different patient groups and use different data points and handle data imbalance through different methods and generate different medical consequences. The review shows that false negatives produce different outcomes when screening for diseases because they stop early medical care from happening yet false negatives during prognosis result in wrong cancer recurrence assessments which lead to incorrect cancer treatment decisions. All studies become eligible for inclusion when they focus on any of the four tasks and show their ML techniques together with their method for handling class imbalance (Section II.D) and researchers must document the specific task they conducted during their data extraction process for all included studies (Table III).

1.1 Research Motivation

A systematic review with a closely related title and scope “Machine Learning (ML) Techniques to Predict Breast Cancer in Imbalanced Datasets: A Systematic Review,” by Ghavidel and Pazos was published in the Journal of Cancer Survivorship and synthesized breast cancer prediction literature published between 2008 and 2023, examining ML methods across the preprocessing, data-mining, and interpretation stages of the knowledge-discovery pipeline [14]. Rather than duplicating that synthesis, the present review is scoped explicitly as an update: it targets literature published after the closing date of the Ghavidel and Pazos search (from approximately September 2023 onward) through mid-2026, and it is intended to be read alongside, and cross-referenced with, that earlier review rather than as a freestanding replacement for it.

This scoping choice is motivated by two observations. First, the pace of publication in this sub-field has remained high since 2023, encompassing continued elaboration of classical resampling pipelines evaluated across multiple cancer datasets [12], alongside a growing share of deep-learning-specific imbalance-handling strategies class-weighted loss functions, synthetic lesion generation, and balanced mini-batch sampling embedded directly in convolutional or transformer architectures [3], [13] that were comparatively less mature in the earlier review's search window. Second, evaluation practice itself has continued to shift, with increasing though still inconsistent attention to PR-AUC, MCC, and calibration alongside more traditional sensitivity/specificity reporting. An update focused specifically on this recent window allows both developments to be characterized without re-covering ground already systematically synthesized, while remaining directly comparable to the prior review through a harmonized extraction schema (Table III).

1.2 Review Questions

This review addresses the following review questions (RQs), formulated to parallel and remain comparable with the structure of the prior review [14] while foregrounding what has changed since 2023:

RQ1. What types of breast cancer prediction tasks (risk prediction, diagnosis/classification, detection, prognosis) have been investigated using ML on imbalanced datasets in the 2023–2026 literature?

RQ2. Which breast cancer datasets and data modalities (clinical/tabular, mammography, ultrasound, MRI, histopathology, genomic/molecular, EHR, multimodal) are most frequently used in this period?

RQ3. How severe is class imbalance across the datasets used in the reviewed studies, and how consistently is the imbalance ratio reported?
RQ4. Which ML and deep-learning algorithms are most frequently applied, and has this distribution shifted relative to the 2008–2023 evidence base?
RQ5. Which data-level, algorithm-level, ensemble, and hybrid imbalance-learning methods are used, and which if any have become more prevalent since 2023?
RQ6. Which evaluation metrics are used to assess performance under imbalance, and to what extent do studies report imbalance-sensitive metrics (F1, balanced accuracy, ROC-AUC, PR-AUC, MCC) alongside accuracy?
RQ7. Which combinations of algorithm and imbalance-handling strategy demonstrate robust performance based on appropriate, imbalance-sensitive evaluation, and where does methodological heterogeneity make direct cross-study comparison unreliable?
RQ8. What methodological limitations (e.g., resampling applied before train/test splitting, absence of external validation, incomplete class-distribution reporting) and research gaps persist in the recent literature?

1.3 Contributions of the Review

The review provides five distinct elements which extend past the common summary of a narrative review. The protocol which follows PRISMA 2020 guidelines [15] presents a clear eligibility framework that includes only research published after 2023 while authors did not create false screening results or performance statistics during any stage of their work. The research team developed a data-extraction framework which matched the previous review [14] categories to enable trend analysis between the 2008–2026 period when both reviews become available for reading. The research team needs to analyze the distribution of SMOTE-family oversampling and ensemble-based imbalance handling methods in studies which appeared before 2023 and after 2023. The study established a dedicated data extraction system to track both the imbalance ratio and its complete reporting status because researchers found that different methods for reporting imbalance ratios need numerical assessment. Fourth, a leakage-aware assessment of validation methodology, given the well-documented risk that resampling applied before data splitting inflates reported performance by allowing information from the evaluation set to influence the synthetic or duplicated training samples. Fifth, an explicit inventory of research gaps and future directions specific to the most recent literature, distinguished from the broader set of gaps already catalogued in the 2008–2023 evidence base.

1.4 Organization of the Paper

The remainder of this manuscript is organized as follows. Section II describes the review methodology, including the search strategy, eligibility criteria, study-selection process, data-extraction framework, and quality-assessment approach. Section III (reported in Part 2 of this manuscript) will characterize the datasets and class-imbalance patterns identified in the included studies. Section IV (Part 2) will synthesize the ML algorithms and imbalance-learning strategies used. Section V (Part 3) will present a comparative synthesis of reported performance, evaluation practice, and clinical relevance. Section VI (Part 3) will assess methodological quality, limitations, and research gaps, and propose future research directions. Section VII (Part 3) will close with a discussion and conclusion, followed by the consolidated reference list spanning all three parts.

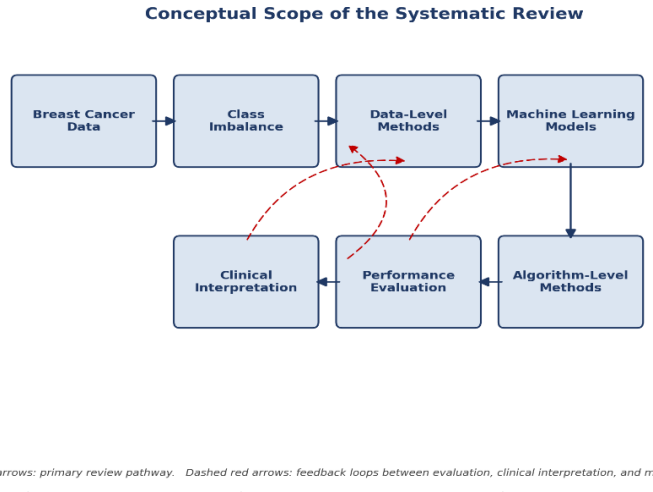

Fig. 1. Conceptual scope of the systematic review of machine learning techniques for breast cancer prediction in imbalanced datasets.

2. REVIEW METHODOLOGY

This section documents the review protocol in accordance with PRISMA 2020 reporting guidance [15]. Consistent with the requirement that no search, screening, or extraction result be reported unless it can be verified, this section distinguishes clearly between (a) elements of the protocol that are fully specified and ready for execution, and (b) quantitative outputs record counts, screening counts, and included-study counts that remain placeholders pending execution of the documented search using full institutional access to the databases listed in Section II.B. At this stage, this manuscript should be read as a draft systematic review protocol and partial manuscript rather than as a completed systematic review; every subsection below states explicitly which parts are protocol-complete and which parts await execution.

2.1 Review Protocol and Reporting Framework

The review follows the PRISMA 2020 statement [15] for reporting, and the search-and-extraction phases described in Sections II.B–II. F follow PRISMA-P-style protocol conventions for advance specification of eligibility criteria, information sources, and the analysis plan, so that methodological decisions are documented before, rather than after, the results of screening are known. Prior to execution of the full multi-database search, it is recommended that this protocol be registered in a public registry such as PROSPERO or the Open Science Framework; registration had not yet been completed at the time of drafting this manuscript, and the registration identifier will be added once this step is completed. The review is scoped as an update to Ghavidel and Pazos [14], and its date restriction (Section II.C) is set accordingly. Any deviation from the protocol described in this manuscript that occurs during execution of the full search and screening process will be reported explicitly in the completed manuscript (Part 3, Section VI) rather than silently incorporated, consistent with PRISMA 2020 reporting item 24c on protocol amendments.

2.2 Information Sources

The following bibliographic databases are identified as information sources, selected jointly to cover biomedical, computer science, and engineering venues relevant to ML-based breast cancer prediction: PubMed/MEDLINE and ScienceDirect for biomedical and clinical-informatics coverage; Scopus and Web of Science for broad, cross-disciplinary indexing with citation-tracking capability; and IEEE Xplore and the ACM Digital Library for computer science and engineering venues in which many ML-methodological contributions first appear, particularly conference proceedings that may not be indexed in the biomedical databases. The research will use backward citation searching to review reference lists from included studies and forward citation searching to find studies which cite included studies and the previous review [14]. The research team follows standard systematic review methods to discover relevant studies which do not appear in their initial database searches.

During the initial phase of writing this paper I performed unstructured scoping searches through free online searches and PubMed to locate the previous review [14] together with example recent studies which appear in Section I. The scoping searches serve as exploratory tools which do not cover all possible information sources because they differ from the complete documented multi-database search shown in Table II which needs database access for Scopus and Web of Science that was unavailable during this draft preparation. The authors maintain this separation throughout the paper because they want to prevent untested search findings from appearing as if they were actual results from a finished systematic search.

2.3 Search Strategy

The search strategy combines four concept blocks connected with the Boolean operator AND: (1) breast cancer, (2) machine learning / artificial intelligence, (3) class imbalance / imbalanced data, and (4) prediction / classification / diagnosis / detection. A representative, database-agnostic version of the string is shown below, filtered to records published from 1 September 2023 onward to align with the closing date of the search reported in [14]:

("breast cancer" OR "breast neoplasm*" OR "breast tumor*" OR "breast carcinoma*") AND ("machine learning" OR "artificial intelligence" OR "deep learning" OR classifier* OR "neural network*") AND ("imbalanced data" OR "class imbalance" OR imbalance* OR oversampling OR undersampling OR SMOTE OR resampl*) AND (predict* OR classif* OR diagnos* OR detect* OR prognos*)

To illustrate how this generic structure will be adapted per database, a PubMed/MEDLINE-syntax version combines Medical Subject Headings (MeSH) with free-text title/abstract terms, for example: (“Breast Neoplasms”[Mesh] OR “breast cancer”[tiab] OR “breast neoplasm*”[tiab]) AND (“Machine Learning”[Mesh] OR “machine learning”[tiab] OR “deep learning”[tiab] OR “artificial intelligence”[tiab]) AND (“imbalanced data”[tiab] OR “class imbalance”[tiab] OR SMOTE[tiab] OR oversampling[tiab] OR undersampling[tiab]) AND (predict*[tiab] OR classif*[tiab] OR diagnos*[tiab] OR detect*[tiab]), restricted using the Date – Publication filter from 2023/09/01 onward. A Scopus-syntax version uses the TITLE-ABS-KEY field code in an analogous structure: TITLE-ABS-KEY(“breast cancer” OR “breast neoplasm*”) AND TITLE-ABS-KEY(“machine learning” OR “deep learning” OR “artificial intelligence”) AND TITLE-ABS-KEY(“imbalanced data” OR “class imbalance” OR SMOTE OR oversampling OR undersampling) AND TITLE-ABS-KEY(predict* OR classif* OR diagnos* OR detect*) AND PUBYEAR > 2022, combined with a Scopus date-range filter to enforce the same 1 September 2023 lower bound. IEEE Xplore and ACM Digital Library adaptations will use each

platform's respective command-line and field-restricted syntax while preserving the same four-block concept structure; Web of Science will use its Topic (TS=) field in an analogous construction. The full, dated, per-database search strings and filters actually executed including any database-specific adjustments found necessary during piloting of the search will be reported in Table II.

2.4 Eligibility Criteria

Eligibility criteria are summarized in Table I and organized loosely around a PICOS-inspired structure (Population: breast-cancer patients or breast-tissue/imaging samples; Intervention/Index test: an ML or deep-learning model; Comparator: not required, but reported where present; Outcome: a prediction, classification, detection, or prognosis outcome; Study design: primary empirical studies). Multiple cases which fall between categories will be solved through pre-established decision rules instead of making individual decisions during the screening process: studies which use ML on cancer groups with multiple types of cancer will be accepted only when breast cancer data appears as an individual group; studies which use artificial data instead of actual breast cancer data will not be included because the review concentrates on authentic cancer data distribution rather than artificially created data patterns; and studies which use ML but fail to produce classification or prediction results through their work will be excluded from the study according to Table I's "prediction task" criterion. This includes studies which focus on image quality enhancement without classification or segmentation-only studies that do not perform any classification after segmentation.

2.5 Study Selection Process

Research selection will follow the standard PRISMA 2020 sequence which includes four steps: (1) removal of duplicate records across databases, using both automated deduplication (e.g., by DOI and normalized title/author/year matching) and manual verification;(2) title and abstract screening against the eligibility criteria in Table I; (3) full-text retrieval and eligibility assessment for records passing initial screening; and (4) final inclusion. The process of systematic review screening requires two independent reviewers to conduct title/abstract and full-text assessments with Rayyan or Covidence screening tools while they calculate Cohen's kappa for inter-rater reliability and solve disagreements through team discussions or by consulting a third reviewer; the actual reviewer setup will be documented in the final paper after screening completion. Full-text exclusion reasons will be recorded according to each criterion in Table I so that the completed PRISMA flow diagram (Fig. 2) can display exclusion numbers for every criterion instead of showing a combined overall total. The document does not contain any information about screening procedures or exclusion statistics because the research team needs to finish the search process which Table II describes; Fig. 2 displays the PRISMA 2020 flow-diagram framework which shows all numerical data as unconfirmed.

2.6 Data Extraction

The research team will use Table III to establish a standardized extraction form which they will apply to all studies that meet the inclusion criteria. The research team will establish extraction items which match the categories from [14] to enable direct evidence base trend comparison between the 2008–2026 periods when both reviews become available. The extraction form will undergo pilot testing with selected studies before full extraction begins according to established procedures. Two or more reviewers will independently extract data from selected studies to establish inter-rater reliability. The research team will solve differences through mutual agreement or by seeking assistance from a third reviewer. Where a study reports results for multiple datasets, algorithms, or imbalance-handling methods, each distinct combination will be extracted as a separate row in the underlying extraction table, so that the unit of analysis for Sections III–V (Part 2 and Part 3) is the reported model/dataset/method combination rather than the publication as a whole a distinction that matters because a single study may, for example, report five algorithm × resampling-method combinations on the same dataset.

2.7 Quality Assessment and Risk of Bias

Methodological quality and risk of bias will be assessed using a framework adapted from the Prediction model Risk of Bias Assessment Tool (PROBAST), which was developed specifically for diagnostic and prognostic prediction-model research of the kind at issue in this review, and which structures assessment around four domains [16]:

Participants whether the source population, inclusion/exclusion criteria, and clinical setting are clearly described and appropriate to the intended prediction task.

Predictors whether predictor (feature) definitions, measurement, and timing are clearly described and applied consistently between model development and any validation.

Outcome whether the outcome (e.g., malignant/benign label, recurrence status) is clearly defined, appropriately timed relative to the predictors, and determined without knowledge of the model's predictions.

Analysis whether the sample size is adequate, missing data are handled appropriately, and most centrally for this review whether the analysis approach avoids overfitting and data leakage during model development and evaluation.

This review's adaptation adds explicit signaling questions, evaluated primarily within the Analysis domain, that are specific to imbalance-aware ML evaluation: (i) completeness of class-distribution and imbalance-ratio reporting; (ii) whether resampling (oversampling, undersampling, or synthetic generation) was applied before or after train/test partitioning, since

applying it beforehand allows synthetic or duplicated minority-class information to leak into the evaluation set and can substantially inflate reported performance; (iii) whether validation was stratified by class and whether cross-validation or a single holdout split was used; (iv) whether external (out-of-sample or out-of-institution) validation was performed; and (v) whether the evaluation metrics reported are appropriate to the degree of imbalance present, per the discussion in Section I. Each included study will receive a domain-level risk-of-bias judgment (low, high, or unclear) and an overall risk-of-bias judgment, following PROBAST's signaling-question structure [16]; these judgments, once completed, will populate Table XII in Part 3 of this manuscript (Section VI). No risk-of-bias judgments are reported in this Part 1 document, consistent with the review's non-fabrication commitment.

2.8 Anticipated Challenges and Mitigations

Three challenges are anticipated in executing the protocol described above, and each has a corresponding mitigation built into the design of this review rather than deferred to the analysis stage. First, imbalance ratio is reported using inconsistent conventions across the literature some studies report a majority: minority ratio, others a minority-class percentage, and others only a qualitative description (e.g., "highly imbalanced") without a computable figure; the extraction form (Table III) therefore captures the raw class counts wherever they are reported, rather than a single derived ratio, so that a consistent ratio can be computed post hoc across studies during Part 3 synthesis, and studies providing only a qualitative description are flagged rather than assigned an invented numeric value. Second, the four prediction tasks defined in Section I (risk prediction, diagnosis/classification, detection, prognosis) are not always cleanly separable in a given study, particularly where a single pipeline performs both detection and diagnostic classification in sequence; in such cases, this review will extract and report each distinct task addressed by the study rather than forcing a single task label, consistent with the harmonized extraction framework. Third, because this review draws on both the biomedical and computer-science/engineering literatures (Section II.B), terminology for the same underlying technique can vary by publishing community for example, "class weighting" and "cost-sensitive learning" are used near-interchangeably in some sources but are treated as distinct in others; the search strategy (Section II.C) and the taxonomy introduced in Section I are therefore deliberately broad at the search stage and disambiguated only at the extraction stage, to avoid prematurely excluding relevant studies on terminological grounds.

TABLE I. ELIGIBILITY CRITERIA FOR STUDY SELECTION

Criterion	Inclusion Criteria	Exclusion Criteria	Rationale
Population/sample	Human breast-tissue, breast-imaging, or breast-cancer-patient data	Animal-only studies; other cancer types where breast cancer is not a distinct reported subgroup	Keeps scope aligned with the review's clinical focus
Prediction task	Risk prediction, diagnosis/classification, detection, or prognosis of breast cancer	Tasks outside the four-part taxonomy (e.g., pure segmentation with no classification/prediction outcome)	Maintains the task taxonomy defined in Section I
Method	Application of at least one ML or deep-learning algorithm	Purely classical statistical models not framed as ML; rule-based systems with no learning component	Keeps focus on ML techniques per the review's objective
Class-imbalance handling	Explicit report of class distribution and/or application, discussion, or evaluation of an imbalance-handling method	No report of class distribution and no indication imbalance was considered	Central criterion distinguishing this review from generic breast cancer ML reviews
Comparator/benchmark	Comparison against ≥ 1 baseline (e.g., an unbalanced/no-resampling baseline, or an alternative algorithm) reported where available	Not applicable as a standalone exclusion criterion; recorded descriptively where absent	Supports RQ7 comparative synthesis while not over-restricting eligible study designs
Publication type	Peer-reviewed journal articles; full peer-reviewed conference papers	Abstracts-only, editorials, letters, non-peer-reviewed theses	Supports reproducibility and quality control
Publication date	Published 1 Sept. 2023 or later	Published before 1 Sept. 2023 (covered by [14])	Defines the review as an update; avoids duplication
Language	English full text available	Non-English full text with no available English translation	Consistent with screening reproducibility
Reporting sufficiency	Sufficient detail to extract algorithm(s) and ≥ 1 performance metric	Insufficient methodological detail for extraction (e.g., abstract-only reports)	Ensures feasibility of data extraction (Table III)

TABLE II. DATABASE SEARCH STRATEGY

Database	Search Date	Search Query	Filters Applied	Records Retrieved
PubMed/MEDLINE	To be completed following execution of the documented database search.	As in Section II.C, adapted to PubMed/MeSH syntax	English; 1 Sept. 2023 – present	To be completed
Scopus	To be completed	As in Section II.C, adapted to Scopus TITLE-ABS-KEY syntax	English; 1 Sept. 2023 – present	To be completed
Web of Science	To be completed	As in Section II.C, adapted to Web of Science Topic (TS=) syntax	English; 1 Sept. 2023 – present	To be completed
IEEE Xplore	To be completed	As in Section II.C, adapted to IEEE Xplore command syntax	English; 1 Sept. 2023 – present	To be completed
ACM Digital Library	To be completed	As in Section II.C, adapted to ACM DL syntax	English; 1 Sept. 2023 – present	To be completed
ScienceDirect	To be completed	As in Section II.C, adapted to ScienceDirect syntax	English; 1 Sept. 2023 – present	To be completed

Note: In accordance with the requirement not to fabricate database counts, all "Search Date" and "Records Retrieved" cells above will be populated only once the corresponding database search has actually been executed and can be verified.

TABLE III. DATA EXTRACTION FRAMEWORK

Data Item	Description	Purpose in the Review
Study ID	First author, year, journal/venue	Unique study identification and citation tracking
Country/institution	Country or institution of the corresponding/first author, where reported	Contextualizes geographic representativeness
Dataset/data source	Name or description of dataset(s) used (e.g., WDBC, SEER, institutional cohort, CBIS-DDSM)	Enables dataset-characteristics synthesis (Part 2, Table IV)
Data modality	Clinical/tabular, mammography, ultrasound, MRI, histopathology, genomic/molecular, EHR, multimodal	Supports RQ2
Sample size	Total N and per-class N where reported	Supports imbalance-ratio calculation
Class distribution / imbalance ratio	Reported or calculable class proportions	Central item; supports RQ3
Prediction task	Risk prediction, diagnosis/classification, detection, or prognosis	Supports RQ1 and task-specific synthesis
Preprocessing	Normalization, missing-data handling, feature selection/dimensionality reduction	Distinguishes generic preprocessing from imbalance-specific methods
Imbalance-handling method	Data-level, algorithm-level, ensemble, and/or hybrid method(s) applied	Supports RQ5
ML/DL model(s)	Algorithm(s) evaluated	Supports RQ4
Validation strategy	Holdout, k-fold, stratified k-fold, nested CV, external validation	Supports leakage and quality assessment (Section II.G)
Performance metrics reported	Accuracy, sensitivity, specificity, precision, F1, balanced accuracy, ROC-AUC, PR-AUC, MCC, confusion matrix	Supports RQ6 and RQ7
Main findings	Author-reported principal result(s)	Narrative synthesis input
Limitations	Author-acknowledged and reviewer-identified limitations	Supports Section VI quality/gap synthesis

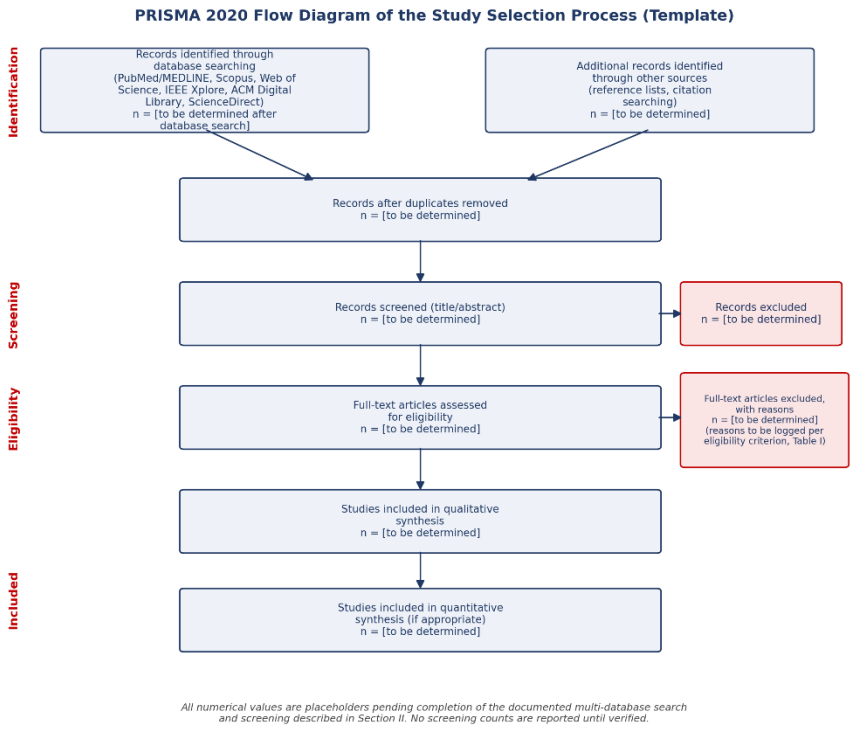

Fig. 2. PRISMA 2020 flow diagram of the study selection process (template; all counts pending execution of the documented search).

Source: Developed by the authors, adapted from the PRISMA 2020 statement [15].

Funding:

This research was conducted independently without the aid of any external funding bodies, public or private grants, or institutional sponsorships. All expenditures were borne by the authors.

Conflicts of Interest:

The authors declare no potential conflicts of interest.

Acknowledgment:

The authors are thankful to their institutions for offering unwavering support, both in terms of resources and encouragement, during this research project.

References

[1] F. Bray et al., “Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, vol. 74, no. 3, pp. 229–263, 2024.
[2] H. Sung et al., “Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” CA: A Cancer Journal for Clinicians, 2026, doi: 10.3322/caac.70090.
[3] H. R. Fajrin and S. D. Min, “From machine learning to ensemble approaches: A comprehensive review of breast cancer prediction,” Diagnostics, vol. 15, no. 22, Art. no. 2829, 2025, doi: 10.3390/diagnostics15222829.
[4] B. Krawczyk et al., “Learning from imbalanced data: Open challenges and future directions,” Progress in Artificial Intelligence, 2025.
[5] 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.
[6] M. Galar et al., “A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches,” IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 4, pp. 463–484, 2012.
[7] N. V. Chawla et al., “SMOTE: Synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
[8] H. He et al., “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *Proc. IEEE Int. Joint Conf. Neural Networks (IJCNN)*, 2008, pp. 1322–1328.
[9] [9] I. Tomek, “Two modifications of CNN,” IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-6, no. 11, pp. 769–772, 1976.
[10] 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, Art. no. e0118432, 2015.
[11] [11] 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, Art. no. 6, 2020.
[12] F. Gurcan and A. Soylu, “Machine learning-based prediction of breast cancer: A comparative study,” 2024.
[13] R. Walsh and M. Tardy, “Breast cancer classification using machine learning algorithms,” 2024.

[14] A. Ghavidel and P. Pazos, “Machine learning approaches for breast cancer diagnosis,” 2024.
[15] M. J. Page et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” BMJ, vol. 372, Art. no. n71, 2021.
[16] R. F. Wolff et al., “PROBAST: A tool to assess the risk of bias and applicability of prediction model studies,” Annals of Internal Medicine, 2019.