Advances in Machine Learning for Tumour Classification in Cancer of Unknown Primary: A Mini-Review

Karen Oróstica, Felipe Mardones, Yanara A. Bernal, Samuel Molina, Marcos Ochard, Ricardo A. Verdugo, Daniel Carvajal-Hausdorf, Katherine Marcelain, Seba Contreras, Ricardo Armisen

PII: S0304-3835(24)00743-2

DOI: https://doi.org/10.1016/j.canlet.2024.217348

Reference: CAN 217348

To appear in: *Cancer Letters*

Received Date: 13 September 2024

Revised Date: 20 November 2024

Accepted Date: 21 November 2024

Advances in Machine Learning for Tumour Classification in Cancer of Unknown Primary: A Mini-Review

Karen Oróstica*[1], Felipe Mardones[1], Yanara A. Bernal[2], Samuel Molina[3], Marcos Ochard[3], Ricardo A. Verdugo[1,5], Daniel Carvajal-Hausdorf[6], Katherine Marcelain[5,7], Seba Contreras*[8], Ricardo Armisen*[2].

[1]Facultad de Medicina, Universidad de Talca, Talca, Chile
[2]Centro de Genética y Genómica, Instituto de Ciencias e Innovación en Medicina, Facultad de Medicina Clínica Alemana Universidad del Desarrollo, Santiago, Chile
[3]Department of Electrical Engineering, Faculty of Physical and Mathematical Sciences, University of Chile, Av. Tupper 2007, Casilla 412-3, Santiago 8370451, Chile
[4]Facultad de Medicina, Universidad de Talca, Talca, Chile
[5]Departamento de Oncología Básico Clínica, Facultad de Medicina, Universidad de Chile, Santiago, Chile
[6]Anatomia Patológica, Clinica Alemana, Facultad de Medicina Universidad del Desarrollo, Santiago, Chile
[7]Centro Para La Prevención y el Control del Cáncer, Universidad de Chile, Santiago, Chile
[8]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany.
* correspondence should be addressed to: korostica09@gmail.com (Karen Oróstica), seba.contreras@ds.mpg.de (Seba Contreras), rarmisen@udd.cl (Ricardo Armisen)

Abstract

Cancers of unknown primary (CUP) are a heterogeneous group of aggressive metastatic cancers where standardised diagnostic techniques fail to identify the organ where it originated, resulting in a poor prognosis and resistance to treatment. Recent advances in large-scale sequencing techniques have enabled the identification of mutational signatures specific to particular tumour subtypes, even from liquid biopsy samples such as blood. This breakthrough paves the way for the development of new cost-effective diagnostic strategies. This mini-review explores recent advancements in Machine Learning (ML) and its application to tumour classification methods for CUP patients, identifying its weaknesses and strengths when classifying the tumour type. In the era of multi-omics, integrating several sources of information (e.g., imaging, molecular biomarkers, and family history) requires important theoretical advancements: increasing the dimensionality of the problem can result in lowering the predictive accuracy and robustness when data is scarce. Here, we review and discuss different architectures and strategies for incorporating cutting-edge machine learning into CUP diagnosis, aiming to bridge the gap between theory and clinical practice.

Keywords
Cancers of unknown primary (CUP); Machine Learning (ML); mutational signatures, somatic mutations; Diagnostic methods; Tumour classification

1. Background

Cancer is a complex group of diseases characterised by abnormal and uncontrolled cell growth with the potential to invade other organs and systems. It arises from the combined effects of multiple genomic and epigenetic factors and is, to date, the second cause of death worldwide, with more than 8 million deaths per year (1). The leading cause of death in cancer patients with solid tumours, metastasis, begins when circulating cancer cells start to colonise distant organs and compromise their function (2,3). Knowing where metastasis originated (i.e., its primary origin) substantially increases the chances of survival, as practitioners have specialised therapeutic alternatives to treat each type of cancer. However, identifying the source is technologically and economically challenging.

Cancers of unknown primary (CUP) are those metastatic cancers where the metastasis's origin is unclear. These cancers are characterised by an aggressive course of the disease and a resistance to conventional chemotherapy, resulting in a poor prognosis for the patients (4). CUP comprises a heterogeneous group of aggressive metastatic tumours with distinct clinicopathological features where standardised diagnostic techniques fail to identify the origin site of cancer at the time of evaluation. This inability to determine the site of origin of the primary tumour represents a significant challenge for modern cancer medicine and society in general, given the technical-economical trade-offs often faced by middle-to-low-income countries in matters of personalised medicine and cutting-edge diagnostic methods (5,6). It is thus of utmost importance to incorporate new multidisciplinary approaches associated with precision medicine that can lead to an earlier, better, and economically accessible identification of the origin.

As the COVID-19 pandemic burdened healthcare systems worldwide and delayed routine controls where cancers could have been detected, we expect the following years to bear disproportionally higher incidences of cancer, particularly of CUP. As diagnostic tools are costly and require high technological investment, there is a prohibitory barrier for vulnerable groups to have precision, genomic and molecular medicine. Therefore, an economically feasible alternative that could help practitioners identify primary tumours in CUP patients will help considerably reduce the associated mortality, especially among least favoured social groups (7). Machine Learning (ML) and Artificial intelligence (AI) methods have made it possible to analyse large volumes of information, identify hidden patterns in the data, and provide increasingly more accurate reports in precision medicine (8). Therefore, integrating AI/ML strategies in the study of CUP patients can result in a substantial leap in the classification of tumour types, the definition of therapeutic regimens, and the prediction of responses to treatment in CUP patients.

This mini-review presents an overview of recent developments in AI/ML-based methods to aid tumour classification and origin identification in CUP patients from genomic profiles of somatic mutations, summarising the epidemiological and methodological aspects of both the disease and methods. We systematically survey the literature to find cutting-edge methods and results that can help overcome economic barriers that currently challenge the application of precision genomic medicine in vulnerable populations, identifying current challenges and research gaps. We aim for this review to serve as an introduction across fields and foster collaboration among multidisciplinary teams developing AI/ML methods in CUP cancer research.

1.1 Epidemiology and subclassification of CUP patients

CUP accounts for 2-5% of all diagnosed cancer cases, being slightly more common in men, with a median age at diagnosis of 65 years (9). However, the percentage can vary depending on the diagnostic capacities of each country. For example, the USA and Asian countries have seen a decline in the incidence of CUP in recent decades, which is attributed to technological advancements in diagnostic methods (10,11). On the other hand, the incidence of CUP is higher among low-income patients who do not have access to an exhaustive diagnostic investigation (12). In other words, heterogeneous diagnostic conditions and capabilities can lead to a premature diagnosis of a patient as CUP, thereby closing the door to targeted therapy.

According to their clinicopathological characteristics, the standard diagnosis classifies patients into two main groups: favourable and unfavourable. The favourable group (15 – 20%) corresponds to patients with a better prognosis, achieving 10 to 16 months survival. By contrast, the unfavourable group (80 – 85%) have a discouraging scenario with few treatment options, presenting a survival of 3 to 6 months (13,14). Although CUP is metastatic cancer by definition, there is a dramatic difference in survival and treatment reception between patients with metastatic cancer with a known primary and CUP patients with a survival of 11.9 and 1.9 months, respectively (14). Therefore, CUP clinically presents unpredictable molecular results correlated with more aggressive clinical parameters and poor survival (15). European ancestry, brain metastases, liver metastases, radiotherapy, and chemotherapy were risk factors for predicting poor overall survival in CUP patients (16).

It is currently possible to know the histological subtype of CUP patients through immunohistochemical (IHC) assays, mucin production, tubule formation, and others (17). Through these assays, the tumours are classified into five known histological groups: well-differentiated adenocarcinomas (12-50%) mainly in females; poorly differentiated adenocarcinomas (15-30%) mainly in younger patients; neuroendocrine carcinoma (46.2%); squamous cell carcinoma (11-15%) mainly in white compared to Asian patients, and undifferentiated neoplasms (5-41%) (11,14,18). However, this histological classification alone can not determine the tissue of origin of the primary tumour (13).

1.2 Advances in Diagnosis and Treatment

CUP diagnosis is a complex process that involves clinical and family history analyses, exhaustive physical examinations, blood tests, biochemical profiles, and tomography and imaging (19). Despite substantial advances in imaging, biomarkers from serum tests and IHC staining, and mRNA expression profiles of biopsy tissues, among other techniques, diagnostic rates of tumours of origin remain low in CUP patients. Currently, CUP patients are treated using multi-agent cytotoxic chemotherapy, given that there are no specific approved drugs to treat CUP patients. Reflecting this lack of specificity, patients' response to treatment is rather poor, with typical rates from 20 to 40% (9). Furthermore, these conventional treatments have not significantly increased overall survival in CUP patients with unfavourable prognoses (17). New treatments based on immunotherapy have been proposed and applied with different degrees of success. In particular, pembrolizumab (anti-PD-1 antibody) is used in patients with CUP tumours. In (20), the authors evaluated the combination of bevacizumab plus erlotinib in CUP patients. However, the response rate was 10%, with an overall survival of 7.4 months. These results were even less encouraging than those obtained by chemotherapy-based treatments.

In a recent prospective, non-randomized, open-label, multicenter Phase II trial (EudraCT 2018-004562-33; NCT04131621), the use of nivolumab and ipilimumab was evaluated in patients with cancer of unknown primary who had relapsed or were refractory after platinum-based chemotherapy. Patients were stratified based on their tumour mutational burden (TMB, high vs. low). Although the study was prematurely terminated before reaching the planned sample size, preliminary results indicated that patients with high TMB (> 12 mutations/Mb) had a higher overall response rate (60% vs. 7.7%) and better median progression-free survival (18.3 vs. 2.4 months) and overall survival (18.3 vs. 3.6 months) compared to those with low TMB. These findings suggest a potential benefit of using TMB as a predictive biomarker in immunotherapy for this type of cancer (21).

Another relevant limitation for studying CUP tumours is the lack of specific gene panels to study the disease and the difficulty of sampling tumour tissue, which reduces the options for molecular diagnostic tests. However, liquid biopsy for tumour DNA has recently gained popularity. In this technique, circulating tumour cells (CTC) and circulating tumour DNA (ctDNA), released early during tumour progression, are used to identify tumour mutations, thus opening a huge window for the genomic characterisation of CUP patients (22). Actionable genetic alterations have been identified in CUP patients by sequencing circulating tumour DNA with a 70-gene panel. The most frequent alterations were present in the TP53, KRAS and PIK3CA genes, all suggesting possible targeted therapies based on their molecular actions (9). Therefore, liquid biopsy could lead to an early diagnosis and simultaneously be the basis for using genomic alterations to choose the best therapeutic strategy.

1.3 Genomic profiles and precision medicine

Somatic mutations observed in cancer can result from numerous mutational processes, with varying intensities and temporality, leaving a characteristic fingerprint on the tumour genome. This fingerprint is known as a mutational signature. Mutational signatures are defined by the type of DNA damage generated and the repair processes involved, resulting in base substitutions, insertions, deletions, or structural variations (23). The advent of large-scale sequencing of genomes and exomes gives us the volume necessary to detect thousands of somatic mutations and thus determine even more mutational signatures than known mutational processes. However, discovering mutational signatures is complex since observable somatic mutations result from multiple overlapping mutational processes (24). However, it has been found that few mutational processes are simultaneously active on cells of a specific type and location within the body. An example is skin cells exposed to the sun and susceptible to ultraviolet radiation (25). Therefore, somatic mutations could be informative of the tissue and the mutational process that originated the mutation, being able to reveal the molecular subtype of cancer by studying these mutational patterns. However, this complex process represents a significant challenge for cancer genomics and bioinformatics since it involves a large volume of information and integrates mathematical methods and sophisticated computational techniques.

One initiative that has addressed this challenge is the Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium. This consortium conducted a comprehensive study of cancer genomes, capturing genetic variation in cancer by sequencing the whole genomes of 2,658 tumours across 38 tumour types. This effort represents a significant advancement in constructing and understanding the cancer genome map, which is highly heterogeneous across multiple levels (cellular, tumour, and patient levels), as it contains the mutational profiles of whole genomes and labelled tumour types. Thanks to these features and the availability of PCAWG, it has been possible to identify clinical-genomic predictors and create multiple tools that integrate bioinformatics, mathematical algorithms, and ML models. However, incorporating these new methodologies in both diagnosis and clinical prognosis has been a challenge due to the complexity of the data and its availability and the multidisciplinary approach required to make sense of it.

The authors in (26) created a cancer-type classifier machine learning model based on the somatic mutational profile obtained after whole-genome sequencing of 2606 tumours representing 24 tumour types as part of the PCAWG Consortium. They achieved a precision of 88% for primary tumour samples and 83% for metastatic samples. Additionally, Mutation-Attention (MuAt), a model based on deep learning, was recently proposed to determine the tumour type based on simple and complex somatic alterations using the same dataset (25). MuAt achieves 89% accuracy, slightly surpassing the model proposed in (26). These works demonstrate that it is possible to use mutational patterns to characterise and distinguish tumours. However, increasing these precision values by including more information in the modelling is still necessary. We must continue to understand the biology of cancer and use it to design cost-effective strategies based on the knowledge we can extract from large volumes of data, such as that generated by whole genome sequencing.

1.4 Machine learning for genetics-based classification in CUP

A classification algorithm in machine learning is a set of logical instructions or rules designed to learn patterns and relationships in labelled data. This type of algorithm belongs to supervised learning, which requires a training dataset that includes examples with known labels. They are used to predict the class or label of a given instance; they take input data and output a categorical variable belonging to a finite set of possible labels. There are two main types of classification algorithms: binary and multiclass. Binary classification algorithms are used when you want to predict between two different classes, for example, whether a patient has cancer or not. Multiclass classification algorithms are used to predict between more than two classes, such as stomach cancer, breast cancer and lung cancer.

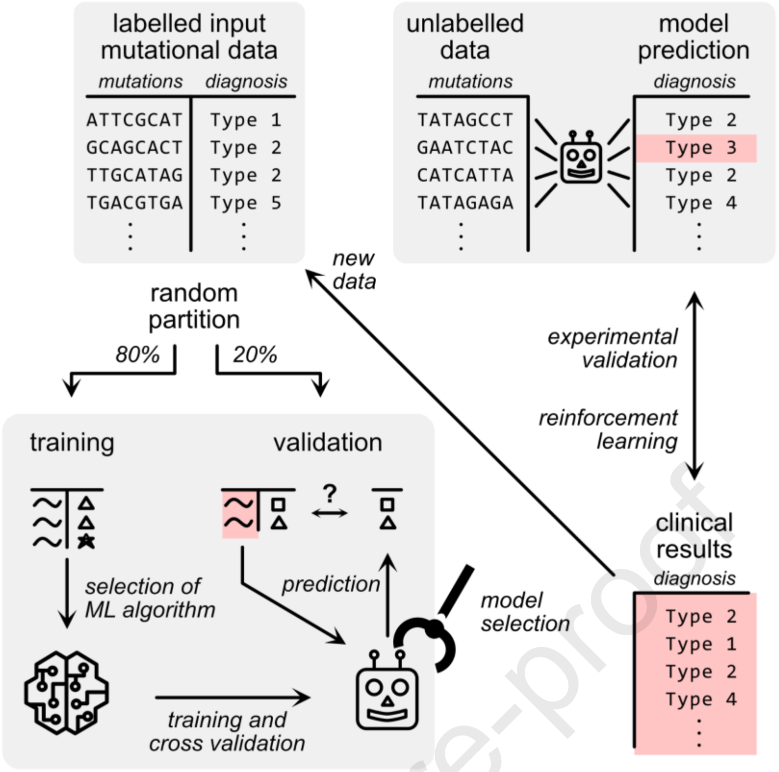

Figure 1. Schematic representation of the methodology to create and validate a classification model using supervised machine learning algorithms. Supervised ML requires the input dataset to have "labels" for each data entry, such as parts of the genome and the type of associated cancer. The input dataset is split randomly into training and validation subsets, usually in an 80:20 ratio. After selecting the ML algorithm, the model is trained using the training dataset, and its parameters are adjusted to fit the data. Then, the model makes predictions using only unlabelled data from the validation dataset. We can compare the predicted labels with the actual (known) labels in the data to assess the model's performance. This process can be repeated multiple times by, for example, creating new random partitions of the input dataset or using different ML models (or hyperparameters) to generate several models. This results in a distribution of performance metrics. Once a model/ensemble is selected based on its performance, it can be used to classify unlabelled data. If some labels are revealed later, they can be added as new data to train new models, and any mismatches can be used to calculate the observed model performance or adapt model parameters dynamically in a reinforcement learning framework.

This mini-review aims to provide an overview of recent advances in ML-assisted tumour classification algorithms based on CUP mutational patterns.

1.5 Challenges and Opportunities for Implementing Machine Learning Models in CUP Classification

The ML models discussed in this review represent major advances in improving tumour classification in CUP patients through genomic data analysis. Given the widespread implementation of genomic techniques in diagnosis and research, we focus on this particular kind of model, but this is only one of the potential data streams in the multi-omics era. Including data such as proteomics, transcriptomics, epigenomics, and family and clinical history in the predictive models offers new opportunities to detect complex, underlying patterns in the data (27). However, this also opens new challenges related to the extra degrees of freedom (high dimensionality) they include. Besides, CUP-specific datasets like The Cancer Genome Atlas (TCGA) and PCAWG remain limited; thus, even complex models will struggle to generalise examples outside of their training data (1,28).

Methodologically, the complex architecture of high-performing models like MuAt challenges the interpretability of the diagnosis, which is crucial in clinical settings. Improving this aspect would encourage practitioners to adopt these technologies more widely (29). On the other hand, ethical and economic implications, especially in low-resource settings, must also be addressed—whether cancer is prematurely classified as CUP is also a matter of how much resources can be invested in diagnosis. Future research should focus on developing cost-effective, computationally efficient models that maintain high accuracy, promoting equitable access to precision medicine for CUP patients globally (30,31). Lastly, clinical validation is essential. While many models perform well retrospectively, prospective validation in clinical trials and integration with current diagnostic workflows are critical for assessing their true impact on patient outcomes (32).

2 Methods

2.1 Search strategy and selection criteria

A systematic literature search was performed using PubMed from 2015 to August 1, 2024. ('cancer of unknown origin'[Title/Abstract] OR 'cancer of unknown primary'[Title/Abstract] OR 'cancer of unknown site'[Title/Abstract] OR 'tissue of origin'[Title/Abstract] OR 'metastatic cancers'[Title/Abstract] OR 'tumo* typ*'[Title/Abstract]) AND ('machine learning'[Title/Abstract] OR predict*[Title/Abstract] OR classif* OR 'deep * learning’) AND (whole[Title/Abstract] OR mutation*[Title/Abstract] OR 'somatic mutation*'[Title/Abstract] OR genetics-based[Title/Abstract] OR mutation patterns[Title/Abstract]). The list of articles was downloaded, duplicates were removed, and articles were excluded if they were focused on the immune system, not in English, did not consider artificial intelligence strategies, did not aim to classify tumour types, or were not related to CUP or were review, systematic review, correction or case report (**Table S1**).

3 Results

We identified 871 potentially eligible studies in the Pubmed database, obtaining n=751 after removing duplicates. Then, we exclude 98 studies by title, abstract and text screening. This left 379 full-text articles that were assessed for eligibility based on five exclusion criteria, 12 of which met the criteria for final inclusion (Figure 2). Performance metrics, model characteristics, and dataset composition are summarised in Table 2. Below, we describe four ML algorithms associated with the largest CUP datasets and their main characteristics.

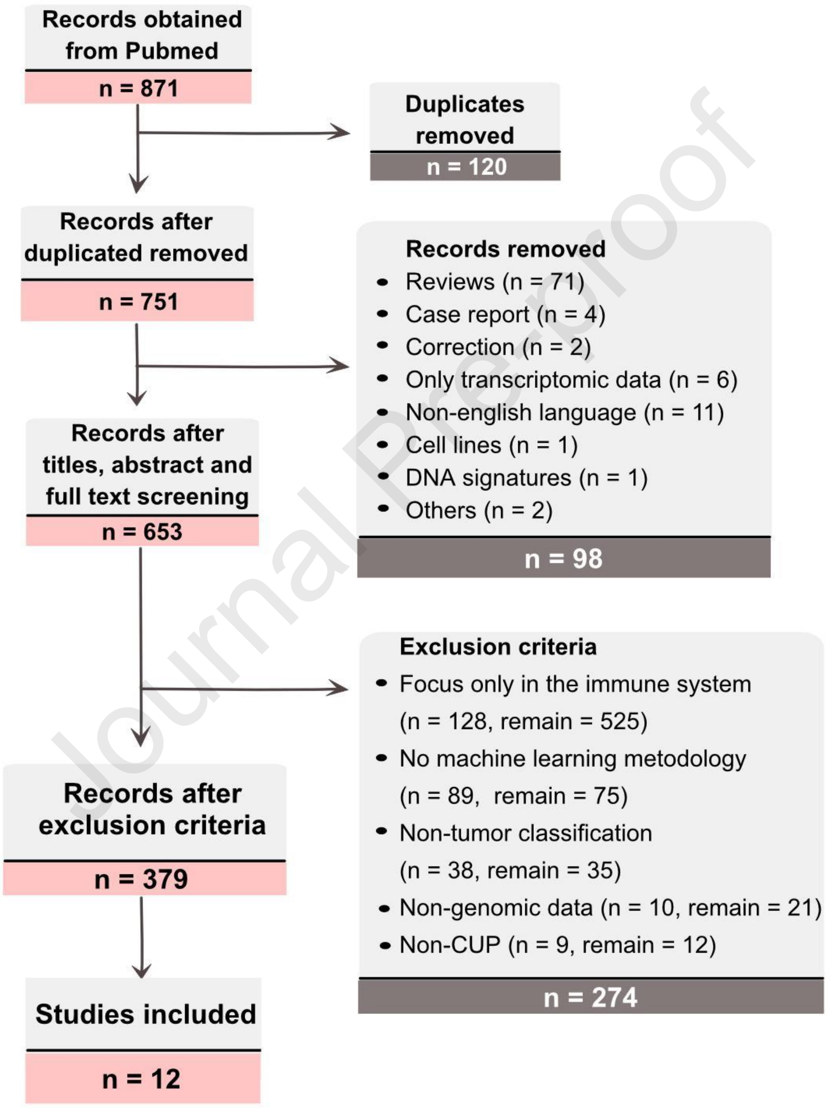
Records obtained from Pubmed
n = 871
Duplicates removed
n = 120
Records after duplicated removed
n = 751
Records removed
Reviews (n = 71)
Case report (n = 4)
Correction (n = 2)
Only transcriptomic data (n = 6)
Non-english language (n = 11)
Cell lines (n = 1)
DNA signatures (n = 1)
Others (n = 2)
n = 98
Records after titles, abstract and full text screening
n = 653
Exclusion criteria
Focus only in the immune system (n = 128, remain = 525)
No machine learning metodology (n = 89, remain = 75)
Non-tumor classification (n = 38, remain = 35)
Non-genomic data (n = 10, remain = 21)
Non-CUP (n = 9, remain = 12)
n = 274
Records after exclusion criteria
n = 379
Studies included
n = 12

Figure 2. Flowchart of search, selection, and inclusion records (adapted from PRISMA (33)).

OncoNPC

The OncoNPC algorithm is an ML classifier developed using a gradient tree boosting framework (XGBoost) to predict cancer types from molecular features. It was trained on targeted next-generation sequencing (NGS) data from 36,445 tumours across 22 cancer types collected in three institutions (34). OncoNPC was trained and validated on 29,176 processed primary and metastasis tumour samples and for training and validation, respectively. Hyperparameter selection was conducted using a random search with 10-fold cross-validation within the training set, using the weighted F1 score as a performance metric. The model was then evaluated on the held-out test set obtaining an overall weighted F1 score of 0.942. To predict the primary sites of CUP tumours, the model was retrained on all CUP tumour samples and applied to the CUP tumours to estimate posterior probabilities across the 22 different cancer labels. For evaluation, OncoNPC was applied to classify 971 CUP tumours from patients admitted to the Dana-Farber Cancer Institute and sequenced as part of routine clinical care, where an F1 value of 0.769 was obtained, being lower than that obtained with known primary tumours.

CUPLR

CUPLR (Cancer of Unknown Primary Location Resolver) (35) integrates WGS-based mutation features, including complex structural variant (SV) features for tumour tissue classification. A harmonised dataset of tumours from 6,756 patients with 35 different cancer types was constructed from two large WGS databases: Hartwig Medical Foundation (Hartwig) and PanCancer Analysis of Whole Genomes (PCAWG). CUPLR extracts around 4,000 features in ten categories, developing a classifier with two main components: binary random forest classifiers and isotonic regressions for probability calibration. One of the key feature categories is regional mutation density (RMD), where 3,071 genomic bins are reduced to 46 profiles per cancer type using non-negative matrix factorisation (NMF). A feature selection process then narrows these down to 511 significant features. The classifier was trained with adjustments for class imbalances and evaluated through stratified cross-validation, achieving a recall and accuracy of 90% in cross-validation and 89% in the test set. CUPLR was also applied to 141 CUP cases in the Hartwig dataset, successfully identifying the cancer type in 58% of cases and providing a strong foundation for classifying CUP samples based on mutational patterns.

A deep learning framework to classify CUP and metastatic cancers using passenger mutation patterns (2020)

The authors in (26) explored cancer type prediction in 2,606 tumours representing 24 cancer types from the PCAWG dataset. The trained classifier is a deep neural network model that accepts input mutation type counts and their genomic positions clustered within each tumour. Deep learning is a subfield of machine learning inspired by the architecture and functioning of biological neural systems. These artificial neural networks consist of multiple layers of nodes (neurones) connected in complex patterns with each other. The term "deep" refers to the complexity of the network and the several layers that constitute it, interconnected hierarchically and optimized to process and find complex patterns in vast amounts of data.

The authors generated seven types of features spanning three main categories. These features were initially evaluated independently and then in combination. Specifically, the distribution of mutations was assessed. The genome was divided into approximately 3,000 1 Mbp bins in the autosomes, and features corresponding to the number of somatic mutations per bin normalised to the total number of somatic mutations were created. Mutation rate profiles were generated independently for somatic single nucleotide variants (SNVs), indels, somatic copy number variations (CNVs), and other structural variations (SVs). They also assessed the mutation type, generating a set of features representing the normalised frequencies of each potential nucleotide change in the context determined by the flanking bases. In addition, they generated nearly 2,000 features related to the high frequency of alterations that distinguish certain tumours to assess driver genes and pathways.

The overall accuracy was highest when only the topological distribution and mutation type of the SNVs were considered. When they tested this deep learning classifier on tumours not considered in training, the accuracy for the complete set of 24 tumour types was 91%. When they applied the classifier trained on PCAWG samples to an independent validation set of 1,436 complete cancer genomes assembled from a series of published non-PCAWG projects, they achieved an overall accuracy of 88% in 14 of the 24 primary cancer types for which the classifier was initially trained. Finally, to assess the classifier's ability to correctly identify the primary tumour type of a metastatic tumour sample, they also constructed an independent dataset consisting of 2,120 samples across 16 tumour types. They achieved an overall accuracy of 83% in identifying the known primary tumour type (26).

TumorTracer

In 2015, Marquard, A. M. et al. designed TumorTracer, a random forest classifier developed to identify a tumour's primary site from its genomic profile (36). The authors used the number of somatic point mutations in a set of 232 frequently mutated genes in cancer, frequencies of the 96 single nucleotide substitution classes determined by the flanking bases, and copy number profiles. This work used the generated features as input to train ten random forest classifiers, one per primary site, i.e., a binary classifier to distinguish between one site and the rest. Each classifier generates a classification score for each primary site. The highest score among all classifiers determines the final classification for a primary site. This score is calculated as the proportion of trees in the forest that voted in favour of the primary site in question.

For the model that used only point mutations, there were ten primary sites and 4,975 patients. On the other hand, for the model that included copy number profile information, there were six primary sites and 2,820 patients. The excluded data (that was not used for training) achieved an overall classification accuracy of 85% at six primary sites (with copy number) and 69% at ten primary sites (without copy number).

Table 1. Machine learning Models for tumour-type classification based on genomic patterns

Study group	Model	Dataset	Size	Algorithms	Performance measures
Moon, et al., 2023 (34)	OncoNPC	Targeted next-generation sequencing (NGS) data collected at the Dana-Farber Cancer Institute (DFCI) Memorial Sloan Kettering (MSK) Cancer Center Vanderbilt-Ingram Cancer Center (VICC)	n=36,445 samples that included 971 CUP samples	XGBoost	Weighted precision: 0.810 Recall: 0.809 Overall weighted F1 score: 0.942
Sanjaya, et al., 2023 (25)	MuAt	2587 whole cancer genomes (PCAWG) 7352 cancer exomes from TCGA	n=10,361 samples	Random Forest Deep Learning/Neural Network (DNN)	Accuracy: 89% for whole genomes Accuracy: 64% for whole exomes
Zelli, et al., 2023 (37)	Not described	TCGA data from cBioportal	9,927 samples spanning 32 different cancer types	XGBoost	out-of-sample balanced accuracy (BACC):77% Area under the curve (AUC): 97%
Huang Y, et al., 2023 (38)	XC-GeM	FH-FMI CGDB, a US nationwide de-identified retrospective longitudinal cancer database	**Training dataset** 12,060 patients **Validation dataset** 955 patients **Independ**	Support vector machines C4.5 decision trees Logistic regression	AUC of 0.965 and Matthew's correlation coefficient (MCC) of 0.742 in the holdout validation dataset

			ent data set 507 patients	Random forest	
Schipper, et al., 2022 (39)	CUPPA	Hartwig database that consisted of samples with known histopathological-based primary origin Netherlands Cancer Institute	**Training dataset** 4058 patients **Testing dataset** 451 patients **Independent data** 72 CUP patients	A statistical model	**testing dataset** AUC of 0.993
Nguyen, et al., 2022 (35)	CUPLR	Pan-Cancer Analysis of Whole Genomes (PCAWG) Hartwig Medical Foundation projects CPCT, DRUP and WIDE	141 Patients Hartwig: 4.902 metastatic tumours PCAWG: 2.835 patient tumours	Random Forest	Overall Accuracy: 90% Recall: 90%
Jiao, et al., 2020 (26)	Not described	PCAWG dataset Hartwig Medical Foundation (HMF data set)	n=2436 samples for training and testing	Random Forest Deep Learning/Neural Network (DNN)	Overall accuracy: 91%
He, et al., 2020 (40)	Not described	Somatic mutation from the Data Portal data Release 28 of the	4909 samples for 13 types of cancers	Random forest	Average accuracy: 0.8822 F1-score: 0.8886

		International Cancer Genome Consortium (ICGC)			
Penson, et al., 2020 (41)	Not described	Memorial Sloan Kettering–Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) clinical cohort	**Training dataset** 7791 patients **Testing dataset** 11 644 patients	Random Forest	Accuracy: 74.1%
Liu, et al., 2020 (42)	Not described	Somatic mutation data from ICGC database version 28	**Training dataset** 3,219 primary samples **Testing dataset** 155 metastatic samples	Random forest	The average accuracy is 86.71%
Marquard, et al., 2015 (36)	TumorTracer	COSMIC database versions 68 and 70 Non-small cell lung cancer patient cohort study (UCLHRTB 10/H1306/42) A cohort of 91 breast metastases cancer from trials SAFIR01 (NCT01414933) and MOSCATO (NCT01566019) 24 specimens from 9 non-small cell lung cancer (NSCLC)	**Training dataset** 7.769 somatic point mutation data (COSMIC v68) **Validation dataset** 1.669 somatic point mutation data (COSMIC v70) 91 breast metastase s samples 24	Stepwise additive logistic regression Artificial Neural Network Support vector machine (SVM) Random forest	Accuracy of 51–64% 69 % PM classifiers (point mutations)

		patients (UCLHRTB 10/H1306/42)	specimens from 9 NSCLC patients		

4. Conclusions and Future Directions

In this mini-review, we systematically searched the literature to identify the most relevant ML-based methods for classifying tumour types based on mutational patterns to date. Although the prevalence of CUP cancers is generally low, there is a strong association between its prevalence and the economic/technological capabilities of the regions where it is diagnosed. These economic barriers to the well-being of all raise the need for new diagnostic methods that allow practitioners and researchers to extract the most information from their data and maximise the probability of early and correct identification of the origin in CUP patients. Predicting the tumour type using sequencing data contributes significantly to the development of precision medicine. With this, it is possible to come closer to providing a therapeutic alternative to CUP patients and, above all, to those CUP patients of the unfavourable type where rapid therapeutic decision-making can impact the patient's survival.

The development of new diagnostic technologies and predictive tools must be interdisciplinary from conception to deployment, integrating the clinical perspective at all stages. Only in that way will the tools leveraging recent advances in ML and sequencing methods be useful in achieving better diagnosis. It is critical, too, that the clinical members of the cancer research community inform themselves about the recent developments in ML/AI and trust its potential to revolutionise the field by creating more precise diagnostic tools. Opening the door to these new technologies is opening the door to a fairer approach to precision medicine, where the synthesising ability of AI/ML models helps personalise medicine at the population scale. In the era of multi-omics, integrating additional sources of information (e.g., imaging, different biomarkers, and family investigations) requires important theoretical advancements: increasing the dimensionality of the problem (i.e., the degrees of freedom, or all the possible values that the new variables/categories may take) can result in lowering the diagnostic power of the method when data is not enough to overcome the increase in complexity. This presents a trade-off: although multi-source data improves our understanding, it also complicates clinical application, where quick, accurate, and low-cost diagnostics are essential. By focusing on mutational data, ML can surface hidden and complex patterns that precisely classify CUP patients and, thereby, improve the health of all.

Availability of data and materials

Not applicable

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Grammarly in order to enhance clarity and conciseness of the writing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Abbreviations

CUP: Cancers of unknown primary
ML: Machine Learning
AI: Artificial intelligence
TMB: Tumour mutational burden
NGS: targeted next-generation sequencing
CUPLR: Cancer of Unknown Primary Location Resolver
SV: structural variant
Hartwig: Hartwig Medical Foundation
PCAWG: PanCancer Analysis of Whole Genomes
NMF: non-negative matrix factorization

References

1. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature. 2020 Feb;578(7793):82–93.
2. Kolling S, Ventre F, Geuna E, Milan M, Pisacane A, Boccaccio C, et al. “Metastatic Cancer of Unknown Primary” or “Primary Metastatic Cancer”? Front Oncol [Internet]. 2020 [cited 2022 Mar 7];9. Available from: https://www.frontiersin.org/article/10.3389/fonc.2019.01546
3. Massagué J, Obenauf AC. Metastatic colonization by circulating tumour cells. Nature. 2016 Jan 21;529(7586):298–306.
4. Rassy E, Assi T, Pavlidis N. Exploring the biological hallmarks of cancer of unknown primary: where do we stand today? Br J Cancer. 2020 Apr;122(8):1124–32.
5. Bochtler T, Krämer A. Does Cancer of Unknown Primary (CUP) Truly Exist as a Distinct Cancer Entity? Front Oncol [Internet]. 2019 [cited 2022 Jan 21];9. Available from: https://www.frontiersin.org/article/10.3389/fonc.2019.00402
6. Binder C, Matthes KL, Korol D, Rohrmann S, Moch H. Cancer of unknown primary—Epidemiological trends and relevance of comprehensive genomic profiling. Cancer Med. 2018 Jul 17;7(9):4814–24.
7. Hamilton W. Cancer diagnostic delay in the COVID-19 era: what happens next? Lancet Oncol. 2020 Aug 1;21(8):1000–2.
8. Johnson KB, Wei W, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision Medicine, AI, and the Future of Personalized Health Care. Clin Transl Sci. 2021 Jan;14(1):86–93.
9. Kato S, Krishnamurthy N, Banks KC, De P, Williams K, Williams C, et al. Utility of Genomic Analysis In Circulating Tumor DNA from Patients with Carcinoma of Unknown Primary. Cancer Res. 2017 Agosto;77(16):4238–46.
10. Boo YK, Park D, Lim J, Lim HS, Won YJ. Descriptive epidemiology of cancer of unknown primary in South Korea, 1999–2017. Cancer Epidemiol. 2021 Oct 1;74:102000.
11. Bytnar JA, Lin J, Moncur JT, Shriver CD, Zhu K. Cancers of Unknown Primary: A Descriptive Study in the U.S. Military Health System. Mil Med. 2023 Mar 1;188(3–4):e516–23.

12. Mnatsakanyan E, Tung WC, Caine B, Smith-Gagen J. Cancer of unknown primary: time trends in incidence, United States. Cancer Causes Control. 2014 Jun 1;25(6):747–57.
13. Olivier T, Fernandez E, Labidi-Galy I, Dietrich PY, Rodriguez-Bravo V, Baciarello G, et al. Redefining cancer of unknown primary: Is precision medicine really shifting the paradigm? Cancer Treat Rev. 2021 Jun 1;97:102204.
14. Rassy E, Assi T, Pavlidis N. Exploring the biological hallmarks of cancer of unknown primary: where do we stand today? Br J Cancer. 2020 Apr;122(8):1124–32.
15. Ren M, Cai X, Jia L, Bai Q, Zhu X, Hu X, et al. Comprehensive analysis of cancer of unknown primary and recommendation of a histological and immunohistochemical diagnostic strategy from China. BMC Cancer. 2023 Dec 1;23(1):1175.
16. Ren Y, Qian S, Xu G, Cai Z, Zhang N, Wang Z. Predicting survival of patients with bone metastasis of unknown origin. Front Endocrinol [Internet]. 2023 Nov 6 [cited 2024 Jul 30];14. Available from: https://www.frontiersin.org/journals/endocrinology/articles/10.3389/fendo.2023.1193318/full
17. Ross JS, Wang K, Gay L, Otto GA, White E, Iwanik K, et al. Comprehensive Genomic Profiling of Carcinoma of Unknown Primary Site: New Routes to Targeted Therapies. JAMA Oncol. 2015 Abril;1(1):40–9.
18. Krämer A, Bochtler T, Pauli C, Baciarello G, Delorme S, Hemminki K, et al. Cancer of unknown primary: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up☆. Ann Oncol. 2023 Mar 1;34(3):228–46.
19. Fizazi K, Greco FA, Pavlidis N, Daugaard G, Oien K, Pentheroudakis G. Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up†. Ann Oncol. 2015 Sep 1;26:v133–8.
20. Hainsworth JD, Spigel DR, Farley C, Thompson DS, Shipley DL, Greco FA, et al. Phase II trial of bevacizumab and erlotinib in carcinomas of unknown primary site: the Minnie Pearl Cancer Research Network. J Clin Oncol Off J Am Soc Clin Oncol. 2007 May 1;25(13):1747–52.
21. Pouyiourou M, Kraft BN, Wohlfromm T, Stahl M, Kubuschok B, Löffler H, et al. Nivolumab and ipilimumab in recurrent or refractory cancer of unknown primary: a phase II trial. Nat Commun. 2023 Oct 24;14(1):6761.
22. Laprovitera N, Salamon I, Gelsomino F, Porcellini E, Riefolo M, Garonzi M, et al. Genetic Characterization of Cancer of Unknown Primary Using Liquid Biopsy Approaches. Front Cell Dev Biol. 2021 Jun 10;9:666156.
23. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014 Sep;15(9):585–98.
24. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020 Feb;578(7793):94–101.
25. Sanjaya P, Waszak SM, Stegle O, Korbel JO, Pitkanen E. Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping [Internet]. bioRxiv; 2022 [cited 2022 Mar 18]. p. 2022.03.15.483816. Available from: https://www.biorxiv.org/content/10.1101/2022.03.15.483816v1
26. Jiao W, Atwal G, Polak P, Karlic R, Cuppen E, Danyi A, et al. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nat Commun [Internet]. 2020 Feb 5 [cited 2020 Apr 17];11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7002586/
27. Huang S, Chaudhary K, Garmire LX. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. Front Genet. 2017 Jun 16;8:84.
28. The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20.
29. Chua IS, Gaziel-Yablowitz M, Korach ZT, Kehl KL, Levitan NA, Arriaga YE, et al. Artificial intelligence in oncology: Path to implementation. Cancer Med. 2021 May 7;10(12):4138.
30. Shreve JT, Khanani SA, Haddad TC. Artificial Intelligence in Oncology: Current Capabilities, Future

Opportunities, and Ethical Considerations. Am Soc Clin Oncol Educ Book. 2022 Jun 10;(42):842–51.

31. Bertsimas D, Wiberg H. Machine Learning in Oncology: Methods, Applications, and Challenges. JCO Clin Cancer Inform. 2020 Oct;(4):885–94.
32. Macheka S, Ng PY, Ginsburg O, Hope A, Sullivan R, Aggarwal A. Prospective evaluation of artificial intelligence (AI) applications for use in cancer pathways following diagnosis: a systematic review. BMJ Oncol [Internet]. 2024 May 10 [cited 2024 Nov 2];3(1). Available from: https://bmjoncology.bmj.com/content/3/1/e000255
33. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Med. 2009 Jul 21;6(7):e1000097.
34. Moon I, LoPiccolo J, Baca SC, Sholl LM, Kehl KL, Hassett MJ, et al. Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. Nat Med. 2023 Aug;29(8):2057–67.
35. Nguyen L, Van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. Nat Commun. 2022 Jul 11;13:4013.
36. Marquard AM, Birkbak NJ, Thomas CE, Favero F, Krzystanek M, Lefebvre C, et al. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med Genomics. 2015 Oct 1;8(1):58.
37. Zelli V, Manno A, Compagnoni C, Ibraheem RO, Zazzeroni F, Alesse E, et al. Classification of tumor types using XGBoost machine learning model: a vector space transformation of genomic alterations. J Transl Med. 2023 Nov 21;21(1):836.
38. Huang Y, Pfeiffer SM, Zhang Q. Primary tumor type prediction based on US nationwide genomic profiling data in 13,522 patients. Comput Struct Biotechnol J. 2023 Jul 26;21:3865–74.
39. Schipper LJ, Samsom KG, Snaebjornsson P, Battaglia T, Bosch LJW, Lalezari F, et al. Complete genomic characterization in patients with cancer of unknown primary origin in routine diagnostics. ESMO Open. 2022 Dec 1;7(6):100611.
40. He B, Dai C, Lang J, Bing P, Tian G, Wang B, et al. A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. Biochim Biophys Acta BBA - Mol Basis Dis. 2020 Nov 1;1866(11):165916.
41. Penson A, Camacho N, Zheng Y, Varghese AM, Al-Ahmadie H, Razavi P, et al. Development of Genome-Derived Tumor Type Prediction to Inform Clinical Cancer Care. JAMA Oncol. 2020 Jan 1;6(1):84–91.
42. Liu X, Li L, Peng L, Wang B, Lang J, Lu Q, et al. Predicting Cancer Tissue-of-Origin by a Machine Learning Method Using DNA Somatic Mutation Data. Front Genet. 2020 Jul 14;11:674.

Acknowledgements

Marcos Orchard would like to thank the Advanced Center for Electrical and Electronic Engineering, Basal Project FB0008. Yanara A. Bernal received support from Agencia Nacional de Investigación y Desarrollo (ANID)/ BECAS/ DOCTORADO NACIONAL/2021– no 21210439. Seba Contreras received support from the Max-Planck-Society.

Funding

This project received support from ANID, Chile, through projects Fondecyt Iniciación 11241185 (KO) and from the IPIG-ACA-KaOr Fund financed by the Gender Equality Instrument of the University of Talca, InES-Gender Project INGE210025 (ANID)(KO), RA research is funded by Proyecto Anillo en Ciencia y Tecnología ACT210079 and FONDECYT regular 1221436.

The funding sources were not involved in the study design, the collection, analysis, and interpretation of data; in the writing of the report, nor in the decision to submit the article for publication.

Author information

Authors and Affiliations

Karen Oróstica, Felipe Mardones and Ricardo A. Verdugo Facultad de Medicina, Universidad de Talca, Talca, Chile

Yanara A. Bernal and Ricardo Armisen Centro de Genética y Genómica, Instituto de Ciencias E Innovación en Medicina, Facultad de Medicina Clínica Alemana Universidad del Desarrollo, Santiago, Chile

Samuel Molina and Marcos Orchard Department of Electrical Engineering, Faculty of Physical and Mathematical Sciences, University of Chile, Av. Tupper 2007, Casilla 412-3, Santiago 8370451, Chile

Daniel Carvajal-Hausdorf Servicio de Anatomía Patológica Clínica Alemana, Santiago, Chile

Katherine Marcelain Departamento de Oncología Básico Clínica, Facultad de Medicina, Universidad de Chile, Santiago, Chile and Centro Para La Prevención y el Control del Cáncer, Universidad de Chile, Santiago, Chile

Seba Contreras Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

Contributions

Conceptualisation: KYO, SC
Methodology: KYO, SC, FM, YB, RA
Investigation: KYO, SC, FM, YB, RA, RAV, KM, DCH, SM, MO
Resources: KYO, YB, FM
Data curation: KYO, YB, FM
Writing - Original Draft: KYO, SC, FM, YB, RA, RAV, KM, DCH, SM, MO
Writing - Review & Editing: KYO, SC, YB, FM
Visualisation: KYO, SM, SC
Supervision: KYO, SC
Project administration: KYO, SC, RA
Funding acquisition: KYO, RA

Corresponding author

Correspondence to Karen Y. Oróstica.

Ethics declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

RA declares honoraria for conferences, advisory boards, and educational activities from Roche, grants, and support for scientific research from Illumina, Pfizer, Roche & Thermo Fisher Scientific, and honoraria for conferences from Thermo Fisher Scientific, Janssen & Tecnofarma and is an Illumina, Thermo Fisher Scientific, Moderna and Pfizer stock holder.

The other authors declare no competing interests.

Highlights

- Cancers of Unknown Primary (CUP) are a diagnostic and clinical challenge in oncology
- Mutational signatures can differentiate tumour tissue and subtypes even in CUP
- Machine Learning leverages hidden and complex patterns in CUP mutational data
- We review recent advancements in ML applied to CUP diagnosis and classification

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The author is an Editorial Board Member/Editor-in-Chief/Associate Editor/Guest Editor for *[Journal name]* and was not involved in the editorial review or the decision to publish this article.

☒ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

RA declares honoraria for conferences, advisory boards, and educational activities from Roche, grants, and support for scientific research from Illumina, Pfizer, Roche & Thermo Fisher Scientific, and honoraria for conferences from Thermo Fisher Scientific, Janssen & Tecnofarma and is an Illumina, Thermo Fisher Scientific, Moderna and Pfizer stock holder.

The other authors declare no competing interests.