Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 222 (2023) 272–283

Procedia Computer Science

www.elsevier.com/locate/procedia

International Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023)

MutFusVAE: Mutational Fusion Variational Autoencoder for Predicting Primary Sites of Cancer

Zhongrui Chen[a], Md Jobayer[b], Md Rakibul Hasan[b,c], Khandaker Asif Ahmed[d], Md Zakir Hossain[a,c,d,*]

[a]*The Australian National University, Canberra ACT 2600, Australia*
[b]*BRAC University, Dhaka 1212, Bangladesh*
[c]*Curtin University, Bentley WA 6102, Australia*
[d]*Commonwealth Scientific and Industrial Research Organisation, Canberra ACT 2601, Australia*

Abstract

The metastatic propensity of malignant primary tumors is a recurring theme when it comes to the cause of mortality in cancer. Establishing the primary site of a metastatic cancer is a significant but challenging task. There are ~3% of metastatic cancer cases diagnosed as cancer of unknown primary (CUP), and the conventional diagnostic process fails to detect the primary site for 80% of CUP patients. Benefiting from the explosion of the information available from large-scale tumor DNA sequencing projects, it became favorable to predict the cancer primary sites from genomic perspective. The existing methods on the task intensively studied the mutational and oncogenic features with assists of machine learning (ML) and deep learning (DL) techniques, yet lack of development of a model architecture tailored for the mutational features. To address the gap, in this research, we aimed to develop a DL methodology specialized for the mutational data. A mutational fusion variational autoencoder (MutFusVAE) deep architecture is proposed to actualize the idea[1]. We downloaded mutational profiles meeting our criteria of 2,603 tumor samples, which get split into 2,082 training samples and 521 (20%) held-out testing samples, from the International Cancer Genome Consortium (ICGC) database. The proposed methods achieved 94% overall classification accuracy for differentiating among seven primary sites on the held-out testing set. The results show the discriminative power brought by a specialized design of deep models for the mutational data and gain insights to facilitate DL-based genomic diagnostics for cancer from a modeling view.

Keywords: cancer; tumor; mutational feature; autoencoder; deep learning

[1] https://github.com/z-r-chen/MutFusVAE-primary-site-prediction

* Corresponding author.
E-mail address: zakir.hossain@anu.edu.au

1877-0509
10.1016/j.procs.2023.08.166

1. Introduction

Cancer is an umbrella term of the group of more than 100 diseases, which are usually characterized by uncontrolled growth and unstoppable spread of abnormal cells. Such abnormality is introduced by the genetic changes that happen in the nucleotide sequences of the genome of an organism. Cancers can be classified in two ways: by the type of tissue where the cancer originates, and by the primary site, which is the organic location in the body where the cancer first developed. The primary site of a cancer may determine the behavior of the tumor and the most likely symptoms, which prepare the family members and guide the cancer care team for properly taking care of the patient. The cancerous cells may spread to other parts of the body, which can be distant from the original site. This type of cancer is called metastatic cancer. There are extreme cases in which the primary sites cannot be detected, leading to cancer of unknown primary (CUP), which accounts for approximately 3% of advanced cancer cases [7]. They are usually diagnosed when the tumor cells have already metastasized to other organs. Because the evidences of metastasis are found at multiple organ sites, the determination of the primary site turns out to be uncertain. Even with extensive diagnostic procedures using modern pathological and medical imaging techniques, the chance of accurate primary site determination remains low. Even though there is a chance to work out the primary site, the conventional diagnostic process can be costly and time-consuming, which puts a burden on the patient's party and delays some treatment to miss the best stages of applying.

More recent large-scale pan-cancer studies involving whole exome and whole genome sequencing techniques have shown that major types of cancer present different mutational patterns in the DNA sequence of tumor cells [11]. A mutation is defined as a change in the DNA sequence of an organism. There are different types of mutations, including single base substitutions (SBS), small insertions and deletions (INDEL), and copy number variations (CNV). The three aforementioned mutational types are the major types we adopted as predictors in this work. An SBS is also referred to as a "point mutation", which means a single nucleotide is altered in the DNA sequence. An INDEL indicates incorporation or loss of DNA fragments, where an extra sequence of nucleotides is inserted into the DNA, or a sequence of existing nucleotides is deleted from the DNA at a specific genomic location. The CNV, on the other hand, refers to a type of intermediate-scale genome structural variation, in which the number of copies of a specific fragment of DNA varies.

Conventional diagnostic tools, including specific pathological investigations (immunohistochemistry, electron microscopy, and molecular diagnosis) and medical imaging techniques (computed tomography (CT) and positron emission tomography (PET) scanning) are commonly used for primary site diagnosis but are often unable to yield precise detection [16]. Recent studies have discovered the potential of using the mutation data detected from the tumor DNA to predict the primary site(s) of cancer [15, 19, 11, 3], revealing the presence of genomic differences among cancer sites. In these works, the discriminatory features are produced based on different types of mutations to train the classifier. For instance, Marquard et al. [15] and Soh et al. [19] both derived features from SBS and CNV. The more recent work by Jiao et al. [11] incorporated the use of INDEL into the features. In addition, they and Chakraborty et al. [3] both took the regional distribution of mutations into account.

Most of the studies which use mutation data to classify the primary site cancer make use of SBS patterns in some manners, as SBSs are typically considered the most reliable information source among different types of mutations [3]. Researchers Jiao et al. [11], Marquard et al. [15] and Chakraborty et al. [3] also derived the trinucleotide-context SBS signature as one of the primary feature sets. Marquard et al. [15] reported the overall predictive accuracy of 58% on 10 cancer primary sites when only using this feature set, where the experiment was conducted on the held-out testing set of the COSMIC database of version 70 [6]. It is worth mentioning that they also tested the results of only using the SBS subtype frequency without the trinucleotide context, i.e., merely considering the substitution types of the replaced and new nucleotides, which achieved only 48% in the same testing environment. Marquard et al. [15] and Soh et al. [19] followed a similar strategy on the derivation of the CNV features, which is to record the presence/absence of mutated genes. By using the combination of SBS and CNV features, they both obtained improvement in predictive performance than using the SBS features alone, suggesting the embodied distinguishable variance of CNVs across the different cancer types. Jiao et al. [11] incorporated INDELs into their somatic mutation feature sets. In their study, except for the derivation of the SBS signature, the SBSs, INDELs, and CNVs were also respectively counted into a regional distribution of each type of mutation.

Marquard et al. [15] developed a set of random forest (RF) classifiers based on the produced features, each of which devoted to generating a classification score for one of the primary sites, and then the predictions were made with the maximum scores, which indicate the most confident inferences. Soh et al. [19] highlighted the results obtained by the support vector machine algorithm with the recursive feature elimination (SVM-RFE) for gene set selection. Chakraborty et al. [3] first projected the high-dimensional mutational features onto a much lower dimensional subspace by applying a principal component analysis (PCA) of 30 principal components followed by a 3D t-distributed stochastic neighbor embedding (tSNE) method. The dimension-reduced features are then fed into a multinomial logistic regression model for the resulting primary site classifier. Jiao et al. [11] developed both an RF classifier and a vanilla multiplayer perceptron (MLP) for comparing their performance on the task of primary site prediction. By using the passenger mutation patterns, they achieved a decent predictive performance of 91% on the held-out tumor samples. Qin et al. [17] experimented with SVM, RF, and MLP algorithms on protein expression and clinical data to classify cancer stages, groups, and treatment outcomes across 32 different cancer types.

The technique of multidimensional variational autoencoder (VAE) has been adopted for a variety of tasks. For instance, Hira et al. [9] and Zhang et al. [21] both developed a multidimensional VAE-based deep framework for integrating the multi-omics data for the downstream tasks of cancer diagnostics. They both harnessed the power of the multidimensional encoder-decoder architecture to learn a low-dimensional latent representation that integrates the knowledge in the high-dimensional multi-omics data. As multiple sources of omics data are used—such as gene-level CNV features, mRNA expressions, and DNA methylation—the data acquisition is comprehensive and thus costly in practice. Still, being inspired by the integration method, we look forward to a joint feature representation of different aspects from pure mutation data. We propose a novel method that separately processes mutational data with different mutation types as different modalities and then integrates the information. Thus, a conceptually better generalization is obtained, capturing the interdependencies among different mutation types and different characteristics of mutation.

In this work, we aim to explore a systemic method of feature extraction from the mutational profile of tumor samples. This procedure produces mutational features concealing the discriminative patterns that can be used for cancer primary site classification. In addition, we aim to find a computational methodology using modern deep learning techniques to accurately infer the cancer primary site from the produced features, yielding a more precise prediction beyond the conventional clinical methods.

2. Materials & methods

2.1. Data collection & processing

The somatic mutation data used for statistical evaluation and model training/testing were downloaded from the open-access portion of release 28 of the International Cancer Genome Consortium (ICGC)[2]. ICGC provides a pan-cancer database that consists of mutation data of tumor samples from different cancer types, each of which is contributed by one or multiple cancer project(s). Most cancer projects listed in the ICGC release 28 provide the 'simple' mutation data, which include the SBSs and INDELs $\leq$ 200 base-pair (bp), while only part of the projects also provides the CNV data.

For example, a 'simple' table and a CNV table are provided by the BRCA-EU project under its data directory, which respectively contains the SBS/INDEL mutations and the CNVs of their sequenced tumor samples, but the BRCA-KR project only provides the 'simple' mutation data. The tables were programmatically processed using the `Pandas` Python library and saved for further use. Then, each table was separated into numerous sub-tables, each of which contains the mutations within a single sample and is named by the corresponding sample ID with the file extension '.csv'. The "sequencing strategy" column in the raw tables specifies the sequencing technique which has been used to detect the corresponding mutation. Since we intend to produce genomic features from genome-wide mutations, only the rows with 'WGS' (whole genome sequencing) in this column were kept, and others were discarded. The 'WGS' value indicates the mutation data of the sample presented in the table are obtained through whole genome sequencing, which means its entire genome has been sequenced to facilitate the sufficient detection of the mutations.

[2] https://dcc.icgc.org/releases/release_28/

The rows with unavailable or invalid required information were also removed. As we need 'ref' and 'alt' nucleotides for constructing the SBS and INDEL mutational signatures, rows with N/A values at these two columns ('ref' and 'alt') were eliminated. The validity of the values in 'chrom' and 'ref'/'alt' columns were examined, and the invalid rows were removed. The 'chrom' is expected to be in the range of $\{1, 2, 3 \ldots 22, X, Y\}$ where chromosomes $1, 2, 3 \ldots 22$ are the autosomes and X and Y are the two sex chromosomes. The nucleotide should be one of the following: A, C, G, or T.

To make use of the CNV features of samples, we excluded the projects which are not found with the CNV data provided via the data portal. A cut-off threshold was set to 100, and the primary sites with samples less than the cut-off threshold value were not considered for the dataset of the classification task. After being processed, the number of samples of each primary site and the corresponding projects, as well as the percentage contributed to the final dataset, is shown in Table 1.

Table 1: The distribution of primary sites of the samples in the dataset.

Primary site	**# samples (percentage)**	**Project(s)**	**# samples**
Breast	732 (28.12%)	BRCA-EU BRCA-US BRCA-FR	569 91 72
Prostate	565 (21.71%)	PRAD-CA PRAD-UK PRAD-FR PRAD-US	328 193 25 19
Brain	457 (17.56%)	PBCA-DE	457
Esophageal	423 (16.25%)	ESAD-UK	423
Pancreas	268 (10.30%)	PACA-CA	268
Bone	98 (3.76%)	BOCA-FR	98
Colorectal	60 (2.31%)	COAD-US READ-US	44 16
TOTAL	2603		

2.2. Methodology

The different cancer types present different patterns of the mutational processes, which can be captured by their resulting mutational signatures. We adapt the mutational signatures proposed by Alexandrov et al. [1] and Steele et al. [20] to capture the mutational characteristics of the tumor samples in the dataset. For each sample, the mutations of types including SBS, INDEL, and CNV are respectively composed into the corresponding mutational signature and also the regional distribution, then they are concatenated to form an integrated mutational landscape. It provides a portrait of the DNA of a tumor cell from the mutational perspective, thus encoding the characteristics of its intrinsic biology.

2.2.1. Derivation of mutational signature

The mutational signature for each mutation type is presented as a part of the crafted feature vector, encapsulating the biological characteristics of that mutation type. The following paragraphs respectively introduce the algorithms for establishing the mutational signatures for each mutation type. The signatures of three mutation types, SBS-sig, INDEL-sig, and CNV-sig, are illustrated on the left side as parts of the whole input feature vector in Fig. 1.

Algorithm 1 Derivation of the SBS signature for a tumor sample

```
procedure GENERATE_SBS_SIGNATURE(tumor)
    sbs_categories ← [A{C > A}A, A{C > A}C, ..., T{T > G}G, T{T > G}T]   ▷ all the 96 possible SBS types
    sbs_sig ← [0, ..., 0]                                                ▷ an array filled with 96 zeros
    for all sbs in tumor do
        chrom, pos, ref, alt ← PARSE_SBS(sbs)
        seq ← FETCH_GENOMIC_SEQUENCE(chrom)
        f5 ← seq[pos − 1]
        f3 ← seq[pos + 1]
        sbs_type ← f5{ref > alt}f3
        i ← ASSIGN_TYPE_INDEX(sbs_type)                                  ▷ Get the index of the type
        sbs_sig[i] ← sbs_sig[i] + 1
    end for
    sbs_sig ← NORMALIZE(sbs_sig)
    return sbs_sig
end procedure
```

SBS signatures. The SBSs are mainly categorized based on two aspects:

- The changing patterns of the ref and alt nucleotides of the substitution
- The trinucleotide-context of the substitution

For the changing patterns, considering the pyrimidines of the Watson-Crick base pairs, there are only six different possible substitutions: C>{A, G, T} and T>{A, C, G}. We also assess the trinucleotide context of the substitution, which is composed of the 5' and 3' flanking bases, i.e., the two directly adjacent nucleotides to the position of the substitution. As there are 4 possible nucleotides (A, C, G, T) for each flanking base, so there are $6 \times 4 \times 4 = 96$ SBS categories in total.

Each SBS is denoted in the HGVSG format, as shown in (1).

$$\langle chrom \rangle : g.\langle pos \rangle \langle ref \rangle > \langle alt \rangle \tag{1}$$

where 'chrom' denotes the chromosome on which the mutation appears, 'pos' denotes the reference position of the mutation in the genomic sequence of the chromosome, and 'ref' and 'alt' respectively denote the reference and alteration (replaced and substituted) nucleotides. To determine the flanking bases, we fetch the genomic sequences from the assembly file using the `pysam` Python library. First, we fetch the genomic sequence of the chromosome from the assembly file, then obtain the 5' and 3' flanking bases by accessing the two adjacent nucleotides to the position of the substitution. By counting all the SBSs of a sample into a distribution reflecting the relative frequencies of the subtypes, an SBS mutational signature is formed. The pseudocode of the full derivation of SBS signature is listed in the Algorithm 1.

INDEL signatures. INDELs are classified into several categories based on the following aspects:

- For the single base insertion and deletion, the type of the inserted/deleted nucleotide and the length of the mononucleotide repeat segment at which they occur.
- For the longer INDELs, the insertion/deletion length and the number of tandem repeat units.

For the 1-bp INDELs, we first assess the single nucleotide inserted or deleted, which can be a C or a T, where A and G will be respectively categorized as the corresponding complementary bases, i.e., T and C. Thereafter, we sub-categorize them based on the length of the mononucleotide repeat segment at which they occur. The mononucleotide

repeats mean a sequence consisting of repetitive nucleotide bases of a single type. For instance, for a deletion happened at the DNA sequence ...GTCTAGGGGGCGCT..., where the deleted part is denoted by the underscore, as the deleted nucleotide is G and in between the mononucleotide sequence of 5 repetitive Gs, the mutation is categorized into the 1bp-del-C-5mono type. For the longer INDELs, we first categorize based on the length of the inserted/deleted sequence and then subtype further by the number of tandem repeat units. The tandem repeat units are defined as a sequence of two or more DNA bases that is repeated numerous times in a head-to-tail manner on a chromosome. For instance, an insertion at the DNA sequence ...CGATATTGCCAGCAGCAGATCGAATGTC..., where the deleted part is denoted by the underscore, as the length of the deleted sequence is 3 and the deleted part is in between three tandem repeat units of CAG, the mutation is categorized into the 3bp-del-3repeats type. In total, 96 types were allocated for the classification of INDELs, which means a 96-dimensional vector equally sized to the SBS is used to represent the INDEL signature.

CNV signatures. Steele et al. [20] developed a framework of 48-channel CNV mutational signature, which incorporates the characteristics including the total copy number (CN), the loss-of-heterozygosity (LOH) status, and the segment size of a CNV. Unfortunately, the first two attributes are not included or only partially provided by some projects in our dataset. The only reliable information source from the CNV data is the segment size, which is mainly ranged by Steele et al. [20] into the following bins: 0-100kb, 100kb-1Mb, 1Mb–10Mb, 10Mb–40Mb, and >40Mb to explain the different scales of CNVs. In their settings, the bins are within each group of subtypes featured by CN and LOH status to produce a reasonable number of categories, but this only produces a limited number of features in the situation where the segment sizes are the only available data. To still harness the remaining CNV mutational features, we established another approach to building up the CNV signature, with a workaround of increasing the covered ranges by intervals of the segment size in the crafted signature to capture more information from the scales of CNVs. Note that we borrowed the `append` and `sorted` operations from Python and the `goto` statement from low-level programming languages for simplicity. The `cnt` is set to 96 to make three mutation signatures possess the same size as shown in Algorithm 2.

Algorithm 2 Expansion of the covered intervals for CNV signature

```
procedure EXPAND_INTERVALS
    s ← [0, 10^5, 10^6, 10^7, 40^7]          ▷ some bonudaries of the intervals used in [20]
    cnt ← 96                                  ▷ the counter to control the size of the produced siganture vector
    while True do
        prev ← s
        for idx in 0, LEN(prev)−1 do
            low, upp ← prev[idx], prev[idx+1]
            mid ← ⌊(low+upp)/2⌋
            S.APPEND(mid)
            if LEN(S) ≥ cnt then
                goto exit_loop
            end if
        end for
        S ← SORTED(S)
    end while
     exit_loop:
    S.APPEND(∞)
    return S
end procedure
```

2.2.2. *Regional mutational distribution*

As Ciriello et al. [5] suggested, the tumors from similar origins will possess similar topological distribution of mutations. Hence, the densities of different mutation types across the genome can be a potential discriminative perspective.

For each type of mutation, we count the locations of mutations to build the topological density of the corresponding mutational event across the genome, which is divided into N_b, representing the total number of bins, each of which contains ~1 million nucleotides. The number of mutations in each bin is calculated to form the regional distribution of the corresponding mutation type. Since CNVs are structural variants that may affect long fragments of DNA, a CNV can cover multiple genomic bins. All the bins are equally incremented by $\frac{1}{N}$, with the amount normalized to capture the range of genomic areas covered by the CNVs, where N is the number of bins covered by the corresponding CNV. The resulting regional mutational distributions for all mutation types are also presented as part of the feature vector, as shown in Fig. 1 with the discussed mutational signatures.

2.2.3. Deep-learning based feature learning

The incorporation of mutational regional distributions leads to an ultra-high-dimensional feature map for each tumor. To calculate the regional distribution of the mutations, the genome is divided into numerous bins. Since we want to derive the regional distribution for each mutation type, the resulting dimension will be equal to $N_b \times N_m$, where N_b is the number of bins and N_m is the number of mutation types. Thus, the vector featuring the mutational landscape will have $N_s+N_m \times N_b$ features, where N_s is the total number of features derived by the mutational signatures. Concretely, this number is calculated as $288+3\times3113 = 9627$, which will lead to the curse of dimensionality problem, given the number of features already exceeds the total count of available samples in the dataset [8]. Besides, the disproportionality of the features is another issue. The hand-crafted features of mutational signatures only take up the minimal portions of the feature vector in comparison to the regional distributions, which may make the discriminative patterns hidden in the signatures overwhelmed by the dominantly sized parts, thus imposing a computational burden to learn the predictive functions. Hence, the dimensionality reduction technique is used to address the problems.

2.2.4. Overall architecture

Being inspired by the variational autoencoder (VAE) framework and multimodal deep learning model, we proposed a mutational fusion variational autoencoder (MutFusVAE) architecture for deciphering the mutational landscape in a multimodal view. We split the feature vector containing the mutational landscape into several parts at the input layer, where each part consists of the features of the mutational signature (Sig) and the regional distribution (RD) to the corresponding mutation type. A group of encoders is used first to obtain a fused representation between the mutational signature and the regional distribution, and then the variational component at the bottleneck encodes the integrated mutational representations to produce a totally fused latent representation, which serves as the compact informative mutational landscape learned by the model. In sum, the hierarchical architecture facilitates representation learning by using a multimodal fashion to integrate the knowledge of interrelationships among features of different mutation types and by cascading the learning task into layered encoder-decoder components.

The model architecture is illustrated in the Fig. 1, where X denotes the feature vector representing the mutational landscape of a tumor sample, which is aimed to be reconstructed as X' at the minimum loss via the MutFusVAE. e_{f_i} and d_{f_i} respectively denote the encoder and decoder for the fused representation between the mutational signature and the regional distribution to the i-th mutation type. e_v and d_v denote the VAE at the bottleneck, which produces a latent representation via encoding the fused features.

Each encoder/decoder block is actualized by single or multiple fully connected layer(s), followed by an activation layer and a batch normalization (BN) layer to address the issue of internal covariate shift [10], except for e_v and d_v, which are not activated and not followed by a BN layer. For the activation function, Sigmoid is used in d_{f_i} to maintain values in the range of $[0, 1]$ for using the binary cross entropy (BCE) function to calculate the reconstruction loss, whilst the rectifier linear unit (ReLU) is used for other layers.

2.2.5. Training strategy

The model training is completed in two stages. The first stage is pre-training the MutFusVAE on the training set in an unsupervised fashion. Then, we replace the decoder parts with a fully connected layer with the output size aligned to the number of predictable classes to form a classification model, which produces the probabilities for final site predictions, and use the labels to supervise the training of the classification model. In the second stage, the pre-trained encoders are fine-tuned to be gradually adapted to the label-related domain. The two training modules are both demonstrated in Fig. 1.

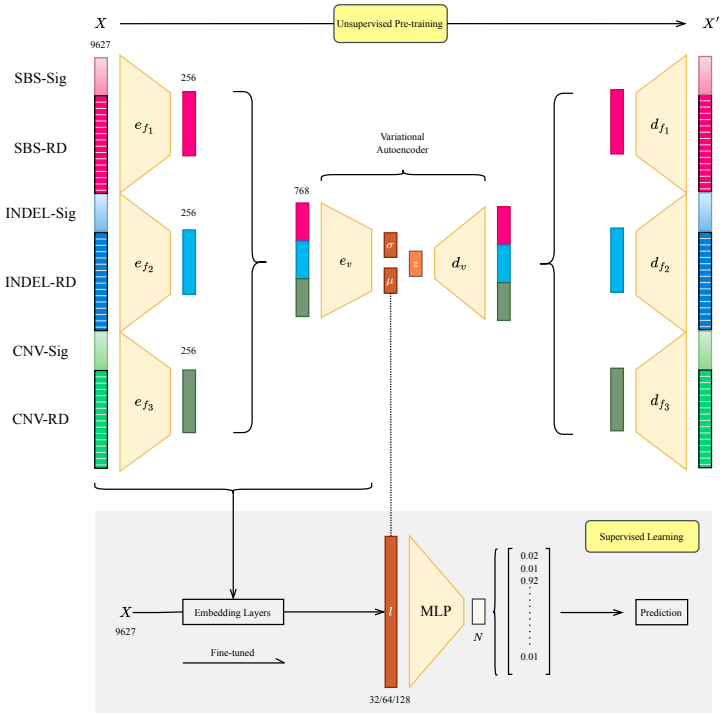


Fig. 1: The overall framework of the proposed mutational fusion variational autoencoder (MutFusVAE), including the modules of unsupervised pre-training and supervised learning. The vector sizes are shown above or below the corresponding vectors, where N is the number of primary sites in the dataset, *Sig* is the mutational signature, and *RD* is the regional distribution.

We applied 10-fold stratified cross-validation to select the best model. We first pre-trained the MutFusVAE using the whole training set without the labels, then randomly split the training set into 10 folds which preserve the same percentage of samples, where each fold is used for validation once.

The pre-training consists of 50 epochs. The supervised training is of 25 epochs, each of which is followed by validation only to save the models which update the current highest balanced accuracy (defined in Sec. 2.2.6). The Adam method [12] of a $1e^{-4}$ learning rate was used for stochastic optimization. The synthetic minority over-sampling technique (SMOTE) [4] sampler was introduced to alleviate the effect brought by the imbalance of the dataset.

2.2.6. *Evaluation metrics*

We evaluated the classification performance in total and for each cancer primary site. The metrics, including overall accuracy, precision, recall, and F_1-score, are defined in (2) to (5), where N is the total number of predictions, y_i stands for i-th primary site; TP, FP, FN stand for true positives, false positives, and false negatives respectively, with respect to the i-th class of primary sites.

$$\text{Accuracy} = \frac{\sum_{y_i} \text{TP}(y_i)}{N} \tag{2}$$

$$\text{Precision}(y_i) = \frac{\text{TP}(y_i)}{\text{TP}(y_i) + \text{FP}(y_i)} \tag{3}$$

$$\text{Recall}(y_i) = \frac{\text{TP}(y_i)}{\text{TP}(y_i) + \text{FN}(y_i)} \tag{4}$$

$$F_1(y_i) = \frac{2 * \text{Precision}(y_i) * \text{Recall}(y_i)}{\text{Precision}(y_i) + \text{Recall}(y_i)} \tag{5}$$

As it is an imbalanced dataset, we also introduce the balanced accuracy (bACC) metric [2], defined in (6), where TPR and TNR are true positives and true negatives normalized by the number of positive and negative samples, respectively.

$$\text{Balanced Accuracy} = \frac{\text{TPR} + \text{TNR}}{2} \tag{6}$$

3. Results & discussions

3.1. Classification performance

To assess the performance of the model, we conducted experiments by doing primary site classification. The training procedure uses 10-fold cross-validation to report the globally averaged metrics and select the best-performing model based on the balanced accuracy to test on the unseen testing set. The held-out testing set has 521 samples, which contribute to 20% of the entire set. The training set and the testing set are generated by a random split in a stratified fashion, and the 10-folds are also obtained by stratified sampling. Thus, the training and validation folds for each split and the testing set all share the same or similar class distribution to prevent the minority classes from being absent in any part, which will significantly affect the results.

The outcomes of final predictions on the held-out testing set are presented in Fig. 2a. The numbers in the diagonal are the numbers of correctly classified test samples to the corresponding primary site. The other cells denote the misclassifications, where the prediction and the corresponding ground truth classes are referred to the row and column. In summary, the final evaluation on the test set yields 94% overall accuracy and 93% balanced accuracy. The testing results and the averaged metrics are summarized in Table 2.

Furthermore, a comparison between our MutFusVAE-based output versus the existing literatures' output has been shown in Table 3.

Table 2: Classification performance reported by 10-fold cross-validation and testing on the held-out testing set. ACC and bACC refer to accuracy and balanced accuracy.

	ACC	**bACC**
CV	0.94 (± 0.02)	0.91 (± 0.03)
Test	0.94	0.93

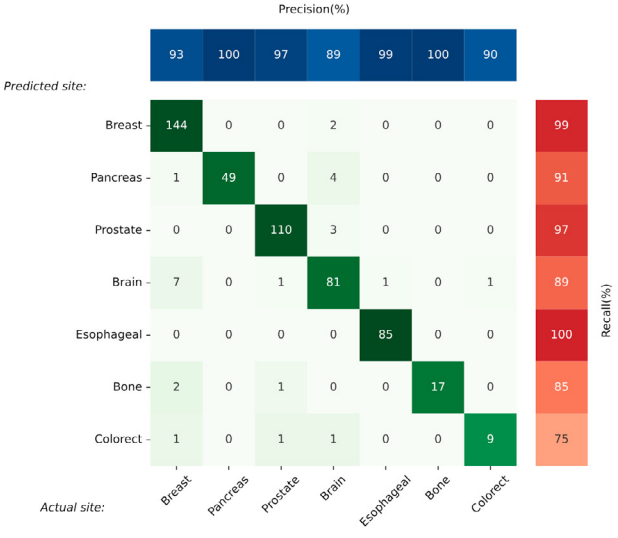
Table 3: Performance comparison among different studies.

Ref	**Best Fit Model**	**# Classes**	**Tumor Samples**	**Accuracy (%)**
Us	DL	7	2603	**94**
[15]	RF	6	1669	85
[19]	Linear SVM	28	6640	77.7
[11]	MLP	24	2606	91

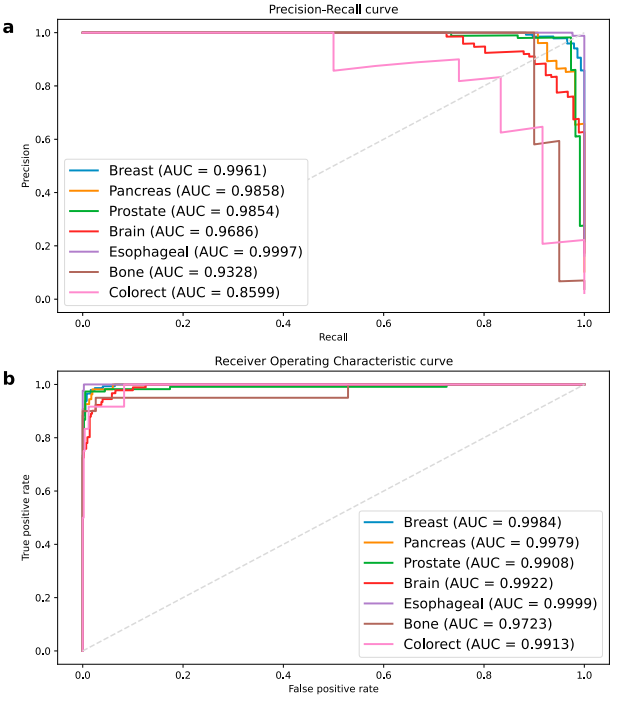
The precision-recall (PR) curve and the receiver operating characteristic (ROC) curve of each primary site versus the rest are displayed in Fig. 2b. The areas under the one-vs-rest PR and ROC curves provide a quantitative summary of the classification performance on differentiating each cancer primary site versus the rest classes. Especially, PR curves can be more informative for imbalanced data than ROC curves, as the former express the susceptibility of classifiers to the imbalanced data [18]. Fig. 2b (b) presents an implication of the strong classification ability of MutFusVAE, as the curves are close to the shape of a perfect classifier, and the AUCs are near to 1. Fig. 2b (a) allows for a more intuitive interpretation of practical classifier performance, with a visual clue that demonstrates the relatively poor performance on the minority class (Colorect).

3.2. Ablation study on feature settings

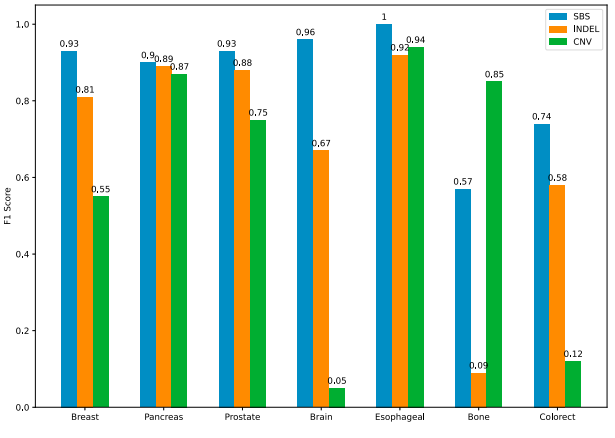
We applied ablation on the feature settings to investigate if they are providing exclusive discriminative information. For each of the SBS, INDEL, and CNV features, we tested the cross-validated F_1 scores of different primary sites. The results shown in Fig. 2c indicate different classes have different levels of specificity and sensitivity to the mutational



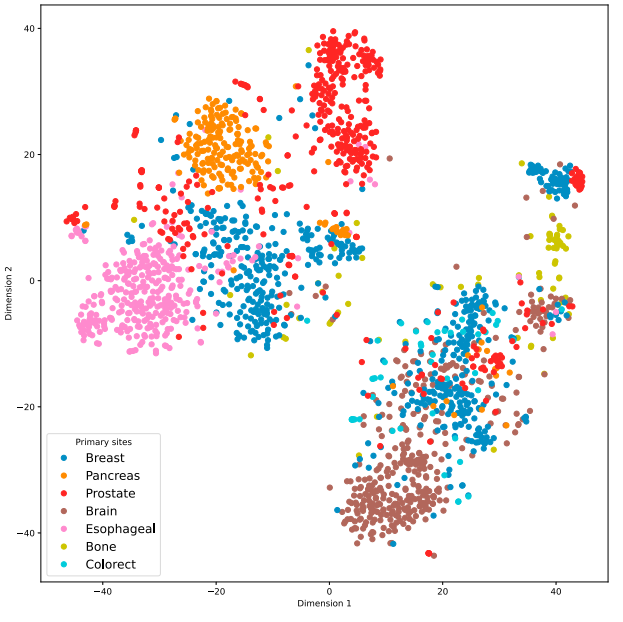
(a) The predictions of the final classifier on the held-out testing set.



(b) 'a' and 'b' show the precision-recall (PR) curve and the receiver operating characteristic (ROC) for each cancer primary site versus the rest. The area under the curve (AUC) is reported in the legend of each graph.



(c) The F_1 scores obtained by using features of every single mutational type.



(d) The 2D representation of the data points obtained by the MutFusVAE via unsupervised training followed by a 2D tSNE. Each point presents a tumor sample in the dataset, which is color-coded corresponding to its primary site.

Fig. 2: The visualization of experimental results.

features of different mutation types. It implies an integrated method will be helpful for discriminating classes that are poorly recognized when only using a single source of mutational features. For example, the incorporation of CNV features is especially helpful for classifying bones.

To inspect the quality of the feature representations learned by the MutFusVAE in the pre-training phase, we extracted its encoding layers and used them to map the high-dimensional features into the latent space. Then, the transformed features were projected to a two-dimensional (2D) subspace using the algorithm of t-distributed stochastic neighbor embedding (tSNE) [14], which is a statistical method widely used for visualizing high-dimensional data. The 2D representation is displayed in Fig. 2d. Each point in the scatter plot represents a tumor sample in the dataset, which is color-coded corresponding to its primary site.

Due to different sources and criteria for data selection, it is hard to compare the performance with the existing studies that conducted experiments on the other dataset or different portions of ICGC. But in our case, a remarkable difference in the proposed methodology compared to other studies is focusing on the pure mutational data extracted from the mutational profiles, while most of the previous studies on the same task more or less harness the oncogenic features. We also attempted to try new model architecture choices rather than the existing classification methods.

Our method is tested with the mutational data, which is developed for mining the underlying mutational patterns, but practical cancer diagnostic incorporates comprehensive procedures. Patients' basic characteristics, such as age, gender, and ethnicity, are also considered valuable information for clinical practitioners and cancer diagnostics. As a result, a full-scale generalizable framework would be extremely beneficial in navigating the power of deep learning techniques on large-scale data integration.

There is also room to improve the transparency of the AI models in use. The deep black box created by layers of artificial neurons is difficult to interpret and thus has varied reliability. Though a certain extent of opacity is inevitable because of the complexity of the algorithm, a more comprehensible methodology should be proposed to unblock the barrier between academia and the real world.

4. Conclusion

Cancer is a hidden illness, making it essential to stop abnormalities and unchecked somatic cell proliferation before they become severe. One of the most important tasks in cancer prevention is the identification of aberrant cells. Traditional detection methods, such as pathology investigations and medical imaging, are exceedingly costly, time-consuming, and insufficiently reliable. So, in this study, we investigated the computational effectiveness of cancer prediction using mutational data with the aid of our proposed MutFusVAE architecture. We pre-processed a number of mutational signatures in order to make them simpler in accordance with the suggested architecture due to an inadequate amount of information available. We interpreted the mutational characteristics in a mutational fusion task setting using the MutFusVAE. Our findings imply that, in the event of a single source of mutational characteristics, an integrated technique will be useful in differentiating poorly recognized classes.

Finally, to implement a full-scale generalized framework, the future scopes involve conducting experiments with large-scale datasets to utilize the power of deep learning techniques on large-scale data integration. Apart from mutational data, analysis of other large-scale data, such as multi-omics data [13], can also be benefited from our MutFusVAE approach. Additionally, improving the transparency of the used AI models would certainly be beneficial, increasing the interpretability of the outcomes. This is more important, especially in biomedical use, because of concerns over ethical issues.

References

[1] Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., Islam, S.M.A., Lopez-Bigas, N., Klimczak, L.J., McPherson, J.R., Morganella, S., Sabarinathan, R., Wheeler, D.A., Mustonen, V., Getz, G., Rozen, S.G., Stratton, M.R., 2020. The repertoire of mutational signatures in human cancer. Nature 578, 94–101. doi:10.1038/s41586-020-1943-3.

[2] Canbek, G., Taskaya Temizel, T., Sagiroglu, S., 2021. BenchMetrics: a systematic benchmarking method for binary classification performance metrics. Neural Computing and Applications 33, 14623–14650. doi:10.1007/s00521-021-06103-6.

[3] Chakraborty, S., Martin, A., Guan, Z., Begg, C.B., Shen, R., 2021. Mining mutation contexts across the cancer genome to map tumor site of origin. Nature Communications 12, 3051. doi:10.1038/s41467-021-23094-z.

[4] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 321–357. doi:10.1613/jair.953.
[5] Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., Sander, C., 2013. Emerging landscape of oncogenic signatures across human cancers. Nature Genetics 45, 1127–1133. doi:10.1038/ng.2762.
[6] Cosmic, . COSMIC - Catalogue of Somatic Mutations in Cancer. URL: https://cancer.sanger.ac.uk/cosmic.
[7] Fizazi, K., Greco, F., Pavlidis, N., Pentheroudakis, G., 2011. Cancers of unknown primary site: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of Oncology 22, vi64–vi68. doi:10.1093/annonc/mdr389.
[8] Friedman, J.H., 1997. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. Data Mining and Knowledge Discovery 1, 55–77. doi:10.1023/A:1009778005914.
[9] Hira, M.T., Razzaque, M.A., Angione, C., Scrivens, J., Sawan, S., Sarker, M., 2021. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. Scientific Reports 11, 6265. doi:10.1038/s41598-021-85285-4.
[10] Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning 37, 448–456. doi:10.48550/ARXIV.1502.03167.
[11] Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Danyi, A., de Ridder, J., van Herpen, C., Lolkema, M.P., Steeghs, N., et al., 2020. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. Nature communications 11, 1–12. doi:10.1038/s41467-019-13825-8.
[12] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .
[13] Liu, X., Hasan, M.R., Ahmed, K.A., Hossain, M.Z., 2023. Machine learning to analyse omic-data for covid-19 diagnosis and prognosis. BMC bioinformatics 24, 1–20. doi:10.1186/s12859-022-05127-6.
[14] van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. Journal of Machine Learning Research 9, 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.
[15] Marquard, A.M., Birkbak, N.J., Thomas, C.E., Favero, F., Krzystanek, M., Lefebvre, C., Ferté, C., Jamal-Hanjani, M., Wilson, G.A., Shafi, S., Swanton, C., André, F., Szallasi, Z., Eklund, A.C., 2015. TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Medical Genomics 8, 58. doi:10.1186/s12920-015-0130-0.
[16] Pavlidis, N., Briasoulis, E., Hainsworth, J., Greco, F., 2003. Diagnostic and therapeutic management of cancer of an unknown primary. European Journal of Cancer 39, 1990–2005. doi:10.1016/S0959-8049(03)00547-1.
[17] Qin, A., Hasan, M.R., Ahmed, K.A., Hossain, M.Z., 2022. Machine learning for predicting cancer severity, in: 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI), pp. 527–529. doi:10.1109/ICHI54592.2022.00098.
[18] Saito, T., Rehmsmeier, M., 2015. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE 10, e0118432. doi:10.1371/journal.pone.0118432.
[19] Soh, K.P., Szczurek, E., Sakoparnig, T., Beerenwinkel, N., 2017. Predicting cancer type from tumour DNA signatures. Genome Medicine 9, 104. doi:10.1186/s13073-017-0493-2.
[20] Steele, C.D., Abbasi, A., Islam, S.M.A., Bowes, A.L., Khandekar, A., Haase, K., Hames-Fathi, S., Ajayi, D., Verfaillie, A., Dhami, P., McLatchie, A., Lechner, M., Light, N., Shlien, A., Malkin, D., Feber, A., Proszek, P., Lesluyes, T., Mertens, F., Flanagan, A.M., Tarabichi, M., Van Loo, P., Alexandrov, L.B., Pillay, N., 2022. Signatures of copy number alterations in human cancer. Nature 606, 984–991. doi:10.1038/s41586-022-04738-6.
[21] Zhang, X., Xing, Y., Sun, K., Guo, Y., 2021. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. Cancers 13, 3047. doi:10.3390/cancers13123047.