Sparse-Input Neural Networks to Differentiate 32 Primary Cancer Types on the Basis of Somatic Point Mutations

Nikolaos Dikaios

Mathematics Research Center, Academy of Athens, 11527 Athens, Greece; ndikaios@academyofathens.gr

Simple Summary: Cancers of unknown primary site represent ~5% of all cancer cases. Most of these cancers receive empirical chemotherapy decided by the oncologist, which typically results in poor survival rates. Identification of the primary cancer site could enable a more rational cancer treatment and even targeted therapies. Given that cancer is considered a genetic disease, one can hypothesize that somatic point mutations could be used to locate the primary cancer type. Studies have shown promising results in identifying breast and colorectal cancer, but there are cancer types/subtypes where somatic point mutations are not performing well. This could be due to somatic point mutations not significantly contributing to cancer initiation but could also be a result of other limitations such as (i) high sparsity in high dimensions, (ii) low signal-to-noise ratio, or (iii) a highly imbalanced dataset. The aim of this research was to examine the ability of somatic point mutations to classify primary cancer types/subtypes from primary tumor samples using state-of-the-art machine learning algorithms.

Abstract: Background and Objective: This paper aimed to differentiate primary cancer types from primary tumor samples on the basis of somatic point mutations (SPMs). Primary cancer site identification is necessary to perform site-specific and potentially targeted treatment. Current methods such as histopathology and lab tests cannot accurately determine cancer origin, which results in empirical patient treatment and poor survival rates. The availability of large deoxyribonucleic acid sequencing datasets has allowed scientists to examine the ability of somatic mutations to classify primary cancer sites. These datasets are highly sparse since most genes will not be mutated, have a low signal-to-noise ratio, and are often imbalanced since rare cancers have fewer samples. Methods: To overcome these limitations a sparse-input neural network (SPINN) is suggested that projects the input data in a lower-dimensional space, where the more informative genes are used for learning. To train and evaluate SPINN, an extensive dataset for SPM was collected from the cancer genome atlas containing 7624 samples spanning 32 cancer types. Different sampling strategies were performed to balance the dataset. SPINN was further validated on an independent ICGC dataset that contained 226 samples spanning four cancer types. Results and Conclusions: SPINN consistently outperformed classification algorithms such as extreme gradient boosting, deep neural networks, and support vector machines, achieving an accuracy up to 73% on independent testing data. Certain primary cancer types/subtypes (e.g., lung, brain, colon, esophagus, skin, and thyroid) were classified with an F-score > 0.80.

Keywords: sparse-input neural network; somatic point mutations; primary cancer site

1. Introduction

The main disciplines used for cancer diagnosis are imaging, histopathology, and lab tests. Imaging is commonly used as a screening tool for cancer and can guide biopsy in hard-to-reach organs to extract tissue samples for histopathological examination. Histopathology can identify cancer cells but cannot always determine the primary site where the tumor originated before metastasizing to different organs. Lab tests usually examine the presence
of proteins and tumor markers for signs of cancer, but the results do not indicate the cancer location and are not conclusive, as noncancerous conditions can cause similar results. Cancer cases of unknown primary site receive empirical treatments and, consequently, have poorer response and survival rate [1]. Given that cancer is a genetic disease, genome analysis could lead to the identification of primary cancer sites and more targeted treatments. Such analysis has recently become feasible due to the availability of large deoxyribonucleic acid (DNA) sequencing datasets.

Cancer type identification using genome analysis involves gene expression signatures, DNA methylation, and genetic aberrations. Gene expression might be the outcome of an altered or unaltered biological process or pathogenic medical condition, which has been used as a predictor of cancer types [2–6]. Abnormal DNA methylation profiles are present in all types of cancer and have also recently been used to identify cancer types [7,8]. This work focuses on a type of genetic aberration, namely, somatic point mutations (SPMs), which play an important role in tumor creation. Spontaneous mutations constantly take place, which accumulate in somatic cells. Most of these mutations are harmless, but others can affect cellular functions. Early mutations can lead to developmental disorders, and progressive accumulation of mutations can cause cancer and aging. Somatic mutations in cancer have been studied more in depth thanks to genome sequencing, providing insight into the mutational processes of genes that drive cancer. Sometimes a mutation can affect a gene or a regulatory element, leading to some cells gaining preferential growth and to clones of these cells surviving. Cancer could be considered as one end-product of somatic cell evolution, which results from the clonal expansion of a single abnormal cell. Martincorena et al. [9] explained how somatic mutations are connected to cancer, although we do not yet have full knowledge of how normal cells become cancer cells.

Somatic point mutations have been used as classifiers of the cancer site [10–14]. The performance of traditional classification algorithms is, however, hindered by imbalances arising from rare cancer types, small sample size, noise, and high data sparsity. Support vector machines (SVMs), classification trees, and k-nearest neighbors perform well for data with complex relations, specifically for low and moderate dimensions, but are not suitable for high-dimensional problems. Neural networks with many layers (deep) according to the circuit complexity theory can efficiently fit complex multivariate functions and perform well on high-dimensional data. Shallower neural networks could in theory perform equally well but would require many hidden units [15]. Deep neural networks (DNNs) require large training datasets and sophisticated stochastic gradient descent algorithms to alleviate the vanishing gradient problem. Most genes, however, do not contain any mutation, which would affect the learning ability of neural networks. Machine learning algorithms such as k-means clustering [14] and inter-class variations [16] have been used to find the discriminatory subset of genes to decrease the complexity of the problem. Identifying a discriminatory subset of genes will not necessarily resolve the problem of sparsity as most of the genes will still not contain a mutation.

To address the issue of sparsity and the lack of high-volume data, various methods have been proposed. Deep Gene [12] used somatic point mutations from 3122 samples and 22,834 genes from The Cancer Genome Atlas (TCGA), which were de-sparsified via two methods called ‘clustered gene filtering’ (CGF) and ‘indexed sparsity reduction’ (ISR) resulting in 1200 features; it achieved an average of 69% accuracy using a DNN classifier. Chen et al. [17] trained an SVM with a linear kernel on >100,000 features extracted from 22,111 genes and 6751 COSMIC samples to classify 17 cancer types. Following a 10-fold cross-validation feature extraction, averages of 62.00% accuracy, 65.24% precision, and 62.26% recall were achieved. Tumor Tracer [18] trained random forest classifiers and through a fivefold cross-validation applied to 530 features consisting of 232 mutations, 232 copy number alterations, and 96 single/trinucleotide base substitution frequencies; it achieved 85.00% accuracy, 85.83% precision, and 84.95% recall over six cancer types and 2820 COSMIC samples. However, when the copy number alterations were excluded, the achieved accuracy dropped to 69%.
This work proposes a sparse-input neural network which ‘directly’ addresses the issue of sparsity using a sparse group lasso regularization. Its performance is validated against commonly used classifiers and extreme gradient boosted trees (XGBoost) [19]. XGBoost is based on the gradient boosting machine; it can represent complex data with correlated features (genes), is robust to noise, and can manage data imbalance. Different balancing strategies were examined as a preprocessing step to examine if their application would benefit the classification accuracy. To evaluate the proposed classifiers, an extensive DNA sequencing database was collected from The Cancer Genome Atlas [20] and the ICGC [21].

2. Materials and Methods

2.1. Theory

Neural networks are not well suited for high-dimensional problems where the number of features \(p\) (e.g., \(p = 22,834\)) is high compared to the number of samples (e.g., \(n = 7624\)). The dataset formulated in this work (described later) is a set of binary features (genes) categorized into 32 cancer types. The formulated database is a case of a multiclass high-dimensional data problem as the number of features \(p\) is high compared to the number of samples. Only 1,974,759 features (genes) from the whole dataset show signs of mutation, whereby around 99% of the data is zero. Highly sparse datasets that contain many zeros (or contain incomplete data with many missing values) pose an additional problem as the learning power decreases due to a lack of informative features. To predict the response of such a complex problem, lasso (least absolute shrinkage and selection operator [22]) terms could be used in the objective function of the neural network to ensure sparsity within each group (cancer type) [23]. The L1 regularization of the neural network first layer weights \(\theta_1\), \(\|\theta_1\|_1\) can result in sparse models with few weights. Consequently, when \(p > n\), it is possible that lasso will tend to choose only one feature out of any cluster of highly correlated feature [24]. More than one gene commonly encodes a cancer type; hence, they should all be included and excluded together. This can be ensured by group lasso [25], which can result in a sparse set of groups; however, all the features in the group will be nonzero. A sparse group lasso penalty suggested by Simon et al., (2013) [26] mixes lasso and group lasso to achieve sparsity of groups and of the features within each group, which better suits the problem at hand. An extension of the sparse group lasso [27] that groups the weights of the first layer to the same input to select a subset of features and uses an additional ridge penalty applied to the weights of all layers other than the first to control their magnitude was used in this work.

\[
\Psi(\theta, \varphi) = \sum_{k=1}^{n} (R_{\theta, \varphi}x_k - y_k)^2 + \lambda_0\|\varphi\|^2_2 + \lambda_1\|\theta_1\|_1 + \lambda_2 \sum_{j=1}^{p} \|\theta(j)\|_2, \tag{1}
\]

where \(R_{\theta, \varphi}\) is the network structure with \(\theta\) the weights of the first input layer and \(\varphi\) the weights of all layers other than the first, \(x\) is the \(p\)-dimensional feature (input) vector, \(y\) is the response variable, and \(\lambda\) represents the regularization parameters. \(x\) is a binary vector of length \(p = 22,834\), where the \(i\)-th component is 0 if the \(i\)-th gene is not mutated and 1 if the \(i\)-th gene is mutated.

2.2. Classifiers

TCGA dataset described in Section 3 (22,834 genes from 7624 different samples spanning 32 cancer types) was split into two sets of samples ensuring the same proportions of class labels as the input dataset: one with 90% training and 10% testing data and the other with 80% training and 20% testing data. Samples were shuffled before splitting and split in a stratified way to ensure the same proportions of class labels between the training and testing datasets. The splitting of the training and testing datasets was repeated 10 times to avoid a misrepresentation of the actual performance of the classifiers due to the particular features of the training and testing datasets in one split. Hyperparameters and/or model parameters were optimized using a grid search approach for each classifier as part of the training. The optimal values were selected on the basis of the best mean cross-
validation accuracy. Tenfold cross-validation was performed for all algorithms. Machine learning algorithms were developed in Python written in Keras [28] with a Tensorflow backend [29]. The developed algorithms were decision tree, k-nearest neighbors, support vector machines, artificial deep neural network, extreme gradient boosting (XGboost), and sparse-input neural nets (SPINNs). The k-nearest neighbors algorithm was run with k = 5. Decision trees were run with maximum depth of the tree equal to 50 and minimum number of samples required to split an internal node equal to 20. Support vector machines were run with regularization parameter C = 0.1 and kernel coefficient gamma = 100. The deep neural network was run with four hidden layers with 8000 neurons each, a total training epoch of 70, a learning rate of (0.001, 0.01, 0.1, 0.2), a weight decay of 0.0005, and a training batch of 256. The ReLu activation function was used, with softmax as the final layer. This a multiclass classification problem where the labels (cancer types) are represented as integers; hence, a sparse categorical cross-entropy objective function was used instead of a categorical cross-entropy. Given the relatively large number of classes (i.e., 32), the softmax function would be quite slow when calculating a categorical cross-entropy. Sparse categorical cross-entropy only uses a portion of all classes, thereby significantly decreasing the computational time. Training was performed by minimizing the sparse categorical cross-entropy, using adaptive learning rate optimization (ADAM [30]). XGBoost is a fast implementation of the gradient boosted decision tree.

Decision trees are regression models in the form of a tree structure. However, they are prone to bias and overfitting. Boosting is a method of sequentially training weak classifiers (decision trees) to produce a strong classifier where each classifier tries to correct its predecessor to prevent bias-related errors. Gradient boosting tries to fit the errors made in the initial fit and correct the corresponding errors in further training. XGBoost was run with a maximum tree depth of 12, boosting learning rate of 0.1, number of boosted trees to fit of 1000, subsampling parameter of 0.9, and sampling level of columns by tree of 0.8. A softmax objective function was used, and multiclass log-loss was used as an evaluation metric. SPINN (described in Section 2.1) was run with three hidden layers (with 2000, 1000, and 500 neurons), a maximum number of iterations of 1000, λ₀ = 0.0003, λ₁ = 0.001, and λ₂ = 0.1. Training was performed by minimizing the objective function in Equation (1) using ADAM [30].

2.3. Sampling Strategies

The two main strategies to deal with imbalanced datasets are either to balance the distribution of classes at the data level or to change the classifier to adapt to imbalanced data at the algorithm level. Data level balancing can be achieved by undersampling, oversampling, or a combination of both. The sampling strategies examined in this work were as follows:

1. Random oversampling, where a subset from minority samples was randomly chosen; these selected samples were replicated and added to the original set.
2. Synthetic minority oversampling technique (SMOTE) [31,32], where oversampling of minority class was achieved by generating synthetic samples.
3. Adaptive synthetic (ADASYN) [33], where the distribution of the minority class was used to adaptively generate synthetic samples.
4. Random undersampling, where data were randomly removed from the majority class to enforce balancing.
5. Tomek link [34], as a modification on CNN (condensed nearest neighbor) [35], where samples were removed from the boundaries of different classes to reduce misclassification.
6. One-sided selection (OSS) [36], where all majority class examples that were at the boundary or noise were removed from the dataset.
7. Edited nearest neighbor (ENN) [37], where samples were removed from the class when the majority of their k nearest neighbors corresponded to a different class.
8. Combination of over- and undersampling, which was performed using SMOTE with Tomek links and SMOTE with edited nearest neighbors.

3. Results
3.1. Reported Dataset

The dataset was collected from TCGA (The Cancer Genome Atlas) [20] with filter criteria IlluminaGA_DNASeq_Curated and was last updated on March 2019. All data can be found at http://tcga-data.nci.nih.gov (accessed on 30 March 2019). This dataset contains information about somatic point mutations in 22,834 genes from 7624 different samples with 32 cancer types. The inclusion of 32 tumor types and subtypes increases the number of associations between tumors and the number of convergent/divergent molecular subtypes. The cancer types are abbreviated as follows: adrenocortical carcinoma (ACC), bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), cholangiocarcinoma (CHOL), colon adenocarcinoma (COAD), lymphoid neoplasm diffuse large b-cell lymphoma (DLBC), esophageal carcinoma (ESCA), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal papillary cell carcinoma (KIRP), acute myeloid leukemia (LAML), brain lower-grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), mesothelioma (MESO), ovarian serous cystadenocarcinoma (OV), pancreatic adenocarcinoma (PAAD), pheochromocytoma and paraganglioma (PCPG), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), sarcoma (SARC), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), testicular germ cell tumors (TGCT), thyroid carcinoma (THCA), thymoma (THYM), uterine corpus endometrial carcinoma (UCEC), uterine carcinosarcoma (UCS), and uveal melanoma (UVM). An overview of the mutations per cancer type (Ca) is shown in Table 1. The number of samples varies heavily between different cancer types (e.g., BRCA has 993 samples whereas CHOL has only 36 samples) making the dataset highly unbalanced.

Table 1. Summary of somatic point mutations per cancer type (Ca).

| Ca       | Sample Number | Missense Mutations | Nonsense Mutations | Nonstop Mutations | Slice Sites | Total Mutations |
|----------|---------------|--------------------|--------------------|-------------------|-------------|-----------------|
| ACC      | 91            | 7258               | 524                | 15                | 415         | 12,082          |
| BLCA     | 130           | 25,108             | 2215               | 47                | 605         | 38,174          |
| BRCA     | 993           | 55,063             | 4841               | 133               | 1561        | 83,973          |
| CESC     | 194           | 26,606             | 2716               | 84                | 574         | 45,936          |
| CHOL     | 36            | 3307               | 316                | 0                 | 90          | 5222            |
| COAD     | 423           | 163,454            | 12,146             | 184               | 9128        | 324,120         |
| DLBC     | 48            | 9623               | 353                | 0                 | 188         | 16,403          |
| ESCA     | 183           | 33,179             | 2829               | 0                 | 851         | 52,576          |
| GBM      | 316           | 26,462             | 2230               | 15                | 813         | 46,899          |
| HNSC     | 279           | 33,260             | 2686               | 44                | 864         | 49,776          |
| KICH     | 66            | 1498               | 103                | 0                 | 54          | 2341            |
| KIRP     | 171           | 9442               | 515                | 18                | 697         | 15,007          |
| LAML     | 197           | 1529               | 117                | 0                 | 188         | 2320            |
| LGG      | 286           | 6098               | 397                | 7                 | 331         | 9256            |
| LIHC     | 373           | 32,553             | 1853               | 0                 | 1238        | 47,895          |
| LUAD     | 230           | 47,700             | 3692               | 56                | 1558        | 69,546          |
| LUSC     | 548           | 162,388            | 12,268             | 301               | 11,708      | 263,481         |
| MESO     | 83            | 2780               | 177                | 8                 | 196         | 5840            |
| OV       | 142           | 7106               | 420                | 9                 | 244         | 10,926          |
| PAAD     | 150           | 19,899             | 1307               | 6                 | 974         | 30,138          |
| PCPG     | 184           | 3271               | 83                 | 0                 | 61          | 4212            |
Table 1. Cont.

| Ca       | Sample Number | Missense Mutations | Nonsense Mutations | Nonstop Mutations | Slice Sites | Total Mutations |
|----------|---------------|--------------------|--------------------|-------------------|-------------|-----------------|
| PRAD     | 332           | 7816               | 433                | 12                | 496         | 11,846          |
| READ     | 81            | 15,899             | 1724               | 0                 | 306         | 23,862          |
| SARC     | 259           | 15,457             | 785                | 19                | 332         | 22,185          |
| SKCM     | 472           | 243,677            | 15,231             | 111               | 10,522      | 423,963         |
| STAD     | 289           | 87,092             | 4423               | 93                | 3852        | 132,196         |
| TGCT     | 156           | 2121               | 123                | 0                 | 52          | 3428            |
| THCA     | 406           | 3000               | 150                | 0                 | 77          | 4489            |
| THYM     | 121           | 15,938             | 1639               | 0                 | 342         | 23,939          |
| UCEC     | 248           | 121,440            | 13,472             | 158               | 1942        | 181,159         |
| UCS      | 57            | 6003               | 612                | 16                | 275         | 8713            |
| UVM      | 80            | 1665               | 73                 | 0                 | 36          | 2856            |
| Total    | 7624          | 1,197,692          | 90,453             | 1336              | 50,436      | 1,974,759       |

The main objectives of the formulated dataset were to compare the performance of different sampling approaches and the proposed machine learning algorithms. To gain better insight into the formulated dataset, intra- and between-class tests were performed on the original dataset before any sampling or splitting was performed. Intraclass correlations were estimated (Table 2) to examine how strong samples in the same cancer class resembled each other. Other than MESO and LAML, the samples of the other cancer types were moderate, good, or excellent. Correspondence analysis was performed to determine the variables response of the gene × sample data in a low-dimensional space. Correspondence analysis can reveal the total picture of the relationship among gene–sample pairs that cannot be performed by pairwise analysis, and it was preferred over other dimension reduction methods because our data consisted of categorical variables. Cumulative inertia was calculated (Figure 1), and it was estimated that 1033 dimensions retained >70% of the total inertia. The intrinsic dimension of the data was also estimated using the maximum-likelihood estimation method [38] and was found to be equal to 811. Both these findings imply overlapping information between different samples.

Table 2. Intraclass correlations for each cancer class.

| ACC     | BLCA    | BRCA    | CESC    | CHOL    | COAD    | DLBC    | ESCA    |
|---------|---------|---------|---------|---------|---------|---------|---------|
| 0.526   | 0.733   | 0.888   | 0.672   | 0.854   | 0.599   | 0.914   | 0.889   |
| GBM     | HNSC    | KICH    | KIRP    | LAML    | LGG     | LIHC    | LUAD    |
| 0.761   | 0.578   | 0.899   | 0.516   | 0.328   | 0.938   | 0.513   | 0.815   |
| LUSC    | MESO    | OV      | PAAD    | PCPG    | PRAD    | READ    | SARC    |
| 0.869   | 0.321   | 0.836   | 0.816   | 0.958   | 0.526   | 0.68    | 0.808   |
| SKCM    | STAD    | TGCT    | THCA    | THYM    | UCEC    | UCS     | UVM     |
| 0.689   | 0.648   | 0.968   | 0.548   | 0.928   | 0.623   | 0.87    | 0.769   |

To further evaluate the performance of the SPINN algorithm on an independent non-TCGA dataset, we used the ICGC (International Cancer Genome Consortium) dataset. Somatic point mutation data were collected for the BRCA, LAML, PAAD, and PRAD primary cancer sites, as shown in Table 3.
### Table 3. Summary of somatic point mutations per cancer type (Ca) for the ICGC dataset.

| Ca.      | Sample Number | Missense Mutations | Nonsense Mutations | Nonstop Mutations | Slice Sites | Total Mutations |
|----------|----------------|--------------------|-------------------|-------------------|-------------|-----------------|
| BRCA     | 60             | 522                | 78                | 0                 | 68          | 668             |
| LAML     | 103            | 752                | 72                | 0                 | 22          | 846             |
| PAAD     | 98             | 154                | 16                | 6                 | 7           | 196             |
| PRAD     | 65             | 328                | 28                | 15                | 1           | 427             |

3.2. Overall Performance of the Classifiers on the Original Dataset

Sparse-input neural networks outperformed the other classifiers on both the 10% and the 20% testing datasets (Table 4). The evaluation was performed using four different metrics, namely, accuracy, precision, recall, and F-score. Accuracy (Acc) is the most commonly used metric measuring the ratio of correctly classified samples over the total number of samples; however, it provides no insights into the ratio of true positives to true negatives. Precision relates to the true positive rate and is equal to the ratio of true positives over the sum of true and false positives. Recall, also referred to as sensitivity, relates to the ratio of correctly classified samples over all samples that have this cancer type, and it is equal to the ratio of true positives over the sum of true positives and false negatives. F-score is more complex to understand but is more reliable than accuracy in our case because the dataset was imbalanced, and the numbers of true positives and true negatives were uneven.

Table 4. Evaluation of the different classifiers on the testing TCGA dataset. The median values (25% to 75% interquartile range) of the metrics are reported over the 10 different splits of the training and testing datasets.

| Learners/Classifiers      | Acc            | Precision        | Recall           | F-Score         |
|---------------------------|----------------|------------------|------------------|-----------------|
|                           | Trained on the 90% of the samples (i.e., 6861) and tested on the 10% of the samples (i.e., 763) |
| Decision Tree             | 0.46 (0.40 to 0.51) | 0.48 (0.42 to 0.51) | 0.38 (0.31 to 0.43) | 0.40 (0.34 to 0.44) |
| KNN                       | 0.44 (0.38 to 0.49) | 0.44 (0.38 to 0.47) | 0.35 (0.30 to 0.39) | 0.33 (0.26 to 0.39) |
| SVM                       | 0.60 (0.55 to 0.64) | 0.64 (0.60 to 0.68) | 0.47 (0.41 to 0.51) | 0.50 (0.44 to 0.53) |
| XGBoost                   | 0.66 (0.62 to 0.68) | 0.64 (0.59 to 0.68) | 0.56 (0.51 to 0.60) | 0.58 (0.53 to 0.63) |
| Neural Networks           | 0.69 (0.64 to 0.73) | 0.66 (0.61 to 0.70) | 0.57 (0.51 to 0.61) | 0.59 (0.54 to 0.63) |
| SPINN                     | 0.71 (0.67 to 0.74) | 0.74 (0.70 to 0.77) | 0.62 (0.57 to 0.66) | 0.65 (0.61 to 0.69) |

![Figure 1. Plot of the cumulative inertia following correspondence analysis.](https://example.com/figure1.png)
Table 4. Cont.

| Learners/Classifiers | Acc          | Precision     | Recall        | F-Score       |
|----------------------|--------------|---------------|---------------|---------------|
|                      | Trained on the 80% of the samples (i.e., 6099) and tested on the 20% of the samples (i.e., 1525) |              |               |               |
| Decision Tree        | 0.45 (0.38 to 0.51) | 0.45 (0.39 to 0.51) | 0.36 (0.29 to 0.41) | 0.38 (0.32 to 0.43) |
| KNN                  | 0.43 (0.35 to 0.49) | 0.45 (0.36 to 0.48) | 0.33 (0.26 to 0.38) | 0.32 (0.25 to 0.38) |
| SVM                  | 0.60 (0.52 to 0.65) | 0.63 (0.56 to 0.68) | 0.47 (0.39 to 0.52) | 0.50 (0.43 to 0.55) |
| XGBoost              | 0.65 (0.59 to 0.70) | 0.63 (0.56 to 0.68) | 0.54 (0.49 to 0.58) | 0.56 (0.50 to 0.60) |
| Neural Networks      | 0.67 (0.60 to 0.72) | 0.63 (0.55 to 0.68) | 0.55 (0.49 to 0.60) | 0.57 (0.50 to 0.61) |
| SPINN                | 0.69 (0.63 to 0.73) | 0.66 (0.61 to 0.71) | 0.59 (0.54 to 0.66) | 0.61 (0.56 to 0.66) |

3.3. Sampling

Initially, the different over/undersampling strategies were applied to the training datasets. For datasets generated after applying oversampling techniques (i.e., SMOTE and ADASYN), the performance of the classifiers remained comparable to the tests on the original datasets. SMOTE gave better results than ADASYN probably due to the samples generated on the outside of the borderline of the minority class. Undersampling methods were also applied to remove many samples from the data. In the case of ENN and CNN, the created dataset contained only 1264 and 772 samples, respectively, from the original data. On the basis of this finding, one could conclude that most of the classes overlapped and had multiple covariates, as also implied by the correspondence analysis (Figure 1). This class overlapping can be considered as the main factor in a classifier’s poor performance besides class imbalance. Due to the reduced number of samples, all classifiers performed poorly on the undersampled data. The only technique that marginally benefited classification (Table 5) was the removal of Tomek links. This approach removes samples from the boundaries of different classes to reduce misclassification.

Table 5. Evaluation of the different classifiers on the testing TCGA dataset (after Tomek links were removed, the total number of samples was reduced to 6859 from 7624). The median values (25% to 75% interquartile range) of the metrics are reported over the 10 different splits of the training and testing datasets.

| Learners/Classifiers | Acc          | Precision     | Recall        | F-Score       |
|----------------------|--------------|---------------|---------------|---------------|
|                      | Trained on the 90% of the samples (i.e., 6173) and tested on the 10% of the samples (i.e., 686) |              |               |               |
| Decision Tree        | 0.46 (0.40 to 0.48) | 0.48 (0.42 to 0.50) | 0.38 (0.33 to 0.40) | 0.40 (0.36 to 0.42) |
| KNN                  | 0.44 (0.39 to 0.46) | 0.44 (0.40 to 0.46) | 0.35 (0.30 to 0.37) | 0.33 (0.29 to 0.36) |
| SVM                  | 0.61 (0.57 to 0.64) | 0.64 (0.60 to 0.67) | 0.47 (0.41 to 0.49) | 0.51 (0.47 to 0.53) |
| XGBoost              | 0.68 (0.63 to 0.71) | 0.65 (0.61 to 0.67) | 0.57 (0.53 to 0.60) | 0.59 (0.56 to 0.61) |
| Neural Networks      | 0.70 (0.65 to 0.73) | 0.65 (0.61 to 0.67) | 0.59 (0.55 to 0.63) | 0.60 (0.55 to 0.62) |
| SPINN                | 0.73 (0.70 to 0.76) | 0.75 (0.72 to 0.78) | 0.64 (0.66 to 0.67) | 0.67 (0.64 to 0.71) |

| Learners/Classifiers | Acc          | Precision     | Recall        | F-Score       |
|----------------------|--------------|---------------|---------------|---------------|
|                      | Trained on the 80% of the samples (i.e., 5487) and tested on the 20% of the samples (i.e., 1372) |              |               |               |
| Decision Tree        | 0.45 (0.39 to 0.50) | 0.45 (0.40 to 0.50) | 0.36 (0.30 to 0.41) | 0.38 (0.33 to 0.42) |
| KNN                  | 0.43 (0.39 to 0.46) | 0.45 (0.40 to 0.49) | 0.33 (0.27 to 0.36) | 0.32 (0.27 to 0.35) |
| SVM                  | 0.60 (0.55 to 0.63) | 0.63 (0.59 to 0.66) | 0.47 (0.42 to 0.50) | 0.50 (0.45 to 0.53) |
| XGBoost              | 0.66 (0.62 to 0.69) | 0.64 (0.60 to 0.67) | 0.55 (0.50 to 0.59) | 0.57 (0.51 to 0.60) |
| Neural Networks      | 0.68 (0.63 to 0.72) | 0.66 (0.61 to 0.70) | 0.57 (0.52 to 0.61) | 0.58 (0.53 to 0.62) |
| SPINN                | 0.71 (0.66 to 0.73) | 0.73 (0.69 to 0.76) | 0.64 (0.60 to 0.67) | 0.66 (0.61 to 0.70) |

3.4. Classifier Performance per Primary Cancer Type

In addition to the overall performance of the classifiers, it is important to examine their performance per cancer type as this varies significantly. Figures 2 and 3 illustrate the performance of the sparse-input neural networks per cancer type (F-score) on the 10% and 20% TCGA testing datasets, respectively. Tables 6 and 7 show the confusion matrices on the 10% and 20% TCGA testing datasets, respectively, to better understand the performance of
the sparse-input neural network. SPINN consistently outperformed the other classification algorithms for all cancer types.

Figure 2. F-score (median value over 10 different splits of training and testing TCGA datasets) per cancer type for the sparse-input neural network on the 10% testing dataset.

Figure 3. F-score (median value over 10 different splits of training and testing TCGA datasets) per cancer type for the sparse-input neural network on the 20% testing dataset.

As expected, the classifier performance varied as a function of cancer type (e.g., 0.24 for OV and 0.94 for LUSC), but this variance should not necessarily be attributed to the sample size. Spearman’s rank correlation coefficient was used to decide whether the sample number and the F-score per cancer type were correlated without assuming them to follow a normal distribution. There was no rank correlation between sample size and F-score ($r = 0.02$ for the 10% testing dataset and $r = 0.04$ for the 20% testing dataset).

The examined classifiers were also validated on an independent non-TCGA dataset (ICGC) that consisted of four primary cancer sites (BRCA, LAML, PAAD, and PRAD). SPINN consistently outperformed the other classifiers (Table 8).

Table 6. Confusion matrix of multiclass classification (columns: predicted, row: true) for the sparse-input neural network on the 10% TCGA testing dataset.
### Table 6. Cont.

| AUC  | BRCA | BCGA | BRCA | CESC | HNSC | KIRP | LGG | LUAD | THCA | THYM | TGCT | UVM | \( L_{AML} \) | \( L_{LUSC} \) | \( L_{SARC} \) | \( L_{UCIC} \) | \( L_{UVM} \) |
|------|------|------|------|------|------|------|-----|------|------|------|------|-----|---------|---------|---------|---------|---------|
| 0.64 | 0.72 | 0.65 | 0.30 | 0.54 | 0.55 | 0.25 | 0.64 | 0.72 | 0.65 | 0.30 | 0.54 | 0.55 | 0.25 | 0.64 | 0.72 | 0.65 | 0.30 |

### Table 7. Confusion matrix of multiclass classification (columns: predicted, row: true) for the sparse-input neural network on the 20% TCGA testing dataset.

| AUC  | BRCA | BCGA | BRCA | CESC | HNSC | KIRP | LGG | LUAD | THCA | THYM | TGCT | UVM | \( L_{AML} \) | \( L_{LUSC} \) | \( L_{SARC} \) | \( L_{UCIC} \) | \( L_{UVM} \) |
|------|------|------|------|------|------|------|-----|------|------|------|------|-----|---------|---------|---------|---------|---------|
| 0.64 | 0.72 | 0.65 | 0.30 | 0.54 | 0.55 | 0.25 | 0.64 | 0.72 | 0.65 | 0.30 | 0.54 | 0.55 | 0.25 | 0.64 | 0.72 | 0.65 | 0.30 |

### Table 8. Evaluation using the F-score of the different classifiers on the independent ICGC dataset. The models validated were trained on TCGA dataset after Tomek links were removed.

| BRCA | LAML | PAAD | PRAD |
|------|------|------|------|
| 0.40 | 0.45 | 0.41 | 0.19 |
| 0.39 | 0.44 | 0.39 | 0.18 |
| 0.53 | 0.55 | 0.54 | 0.25 |
| 0.62 | 0.69 | 0.61 | 0.30 |
| 0.61 | 0.68 | 0.62 | 0.29 |
| 0.64 | 0.72 | 0.65 | 0.30 |
4. Discussion

Cancers of unknown primary site represent ~5% of all cancer cases. Most of these cancers receive empirical chemotherapy decided by the oncologist, which typically results in poor survival rates. Identification of the primary cancer site could enable a more rational cancer treatment and even targeted therapies. Given that cancer is considered a genetic disease [39], one can hypothesize that somatic point mutations could be used to locate the primary cancer type. Studies have shown promising results on identifying breast and colorectal cancer [39], but there are cancer types/subtypes where somatic point mutations are not performing well. This could be due to somatic point mutations not significantly contributing to cancer initiation but could also be a result of other limitations such as (i) high sparsity in high dimensions, (ii) low signal-to-noise ratio, or (iii) a highly imbalanced dataset. With next-generation cost-effective gene sequencing, we are getting a high amount of genomics data. The aim of this research was to examine the ability of somatic point mutations to classify primary cancer types/subtypes from primary tumor samples using state-of-the-art machine learning algorithms.

TCGA open-access data were collected as described in Section 2, which consisted of 22,834 genes from 7624 different samples spanning 32 different cancer types. To the best of the author’s knowledge, this is the first time such an extensive dataset with samples from 32 cancer types has been reported. The resulting database is very imbalanced with common cancers sites such as breast having 993 samples, while rare cancer sites have as low as 36 samples. All 22,834 genes were included, resulting in a highly sparse database with 99% of the genes having no mutations. Different machine learning algorithms were trained on 90% or 80% of the original dataset and were tested on the remaining 10% or 20%, respectively. An independent validation was also performed for the BRCA, LAML, PAAD, and PRAD primary cancer sites using samples collected for the ICGC.

Neural networks perform well on high-dimensional problems and can approximate complex multivariate functions; however, given that only a small subset of the genes would be informative per cancer type, their performance was hindered. This work proposed a sparse-input neural network (described in Section 2.1) which employs a combination of lasso, group lasso, and ridge penalties to the loss function to project the input data in a lower-dimensional space where the more informative genes are used for learning. Our results show that the sparse-input neural network could achieve up to 73% accuracy on TCGA dataset without any preprocessing of features such as gene selection. The above statement shows the learning power of neural networks with regularization. XGBoost and deep neural networks also performed well compared to traditional classifiers (decision trees, KNN, and SVM). These findings were also confirmed when the trained classifiers were validated on the ICGC independent dataset.

All sampling strategies described in the literature are associated with the use of nearest neighbors to either oversample or undersample the dataset. In this work, balancing TCGA dataset using sampling strategies did not benefit the classifier performance except for removing Tomek links. This was probably due to the high amount of class overlap. Figures 2 and 3 demonstrate that classification performance significantly varied as a function of cancer type. In agreement with previous studies, breast and colorectal cancer had a high classification accuracy (F-scores up to 0.73 and 0.90, respectively). This study showcased that somatic point mutations can also accurately classify other types of cancer. There were cancer types, however, where classifiers performed poorly. This is not necessarily related solely to having few training samples, as the F-score did not seem to relate to the sample size; however, for certain cancer types, it could also be related to a high amount of class overlap. This hypothesis was reinforced following ENN, CNN undersampling, correspondence analysis, and estimation of the intrinsic dimension. All analyses suggested that only ~1000 of the samples were mutually independent.
5. Conclusions

To conclude, this work determined that using only somatic point mutations can yield good performance in differentiating primary cancer types if the sparsity of the data is considered. Results, however, also indicated some similarity in the information provided by somatic point mutations for different primary cancer types. This limitation could be managed by (i) investigating preprocessing methods [40–42] that could cluster somatic mutations and/or learn which genes are involved in cancer initiation [43], and (ii) enriching the database especially for rare cancer types and/or introducing additional genomic information such as copy number variations, as well as DNA methylation and gene expression signatures.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available on TCGA website.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pavlidis, N.; Pentheroudakis, G. Cancer of unknown primary site. Lancet 2012, 379, 1428–1435. [CrossRef]
2. Liu, J.; Campen, A.; Huang, S.; Peng, S.; Ye, X.; Palakal, M.; Dunker, A.; Xia, Y.; Li, S. Identification of a gene signature in cell cycle pathway for breast cancer prognosis using gene expression profiling data. BMC Med. Genom. 2008, 1, 39. [CrossRef] [PubMed]
3. Golub, T.; Slonim, D.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J.; Coller, H.; Loh, M.L.; Downing, J.R.; Caligiuri, M.A.; et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 1999, 286, 531–537. [CrossRef] [PubMed]
4. Khan, J.; Wei, J.; Ringnér, M.; Saal, L.; Ladanyi, M.; Westermann, F.; Berthold, F.; Schwab, M.; Antonescu, C.R.; Peterson, C.; et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 2001, 7, 673–679. [CrossRef]
5. Ramaswamy, S.; Tamayo, P.; Rifkin, R.; Mukherjee, S.; Yeang, C.; Angelo, M.; Ladd, C.; Reich, M.; Latulippe, E.; Mesirov, J.P.; et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc. Natl. Acad. Sci. USA 2001, 98, 15149–15154. [CrossRef] [PubMed]
6. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA 2002, 99, 6567–6572. [CrossRef]
7. Kang, S.; Li, Q.; Chen, Q.; Zhou, Y.; Park, S.; Lee, G.; Grimes, B.; Krysan, K.; Yu, M.; Wang, W.; et al. CancerLocator: Non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. Genome Biol. 2017, 18, 53. [CrossRef]
8. Hao, X.; Luo, H.; Krawczyk, M.; Wei, W.; Wang, W.; Wang, J.; Flagg, K.; Hou, J.; Zhang, H.; Yi, S.; et al. DNA methylation markers for diagnosis and prognosis of common cancers. Proc. Natl. Acad. Sci. USA 2017, 114, 7414–7419. [CrossRef]
9. Martincorena, I.; Campbell, P. Somatic mutation in cancer and normal cells. Science 2015, 349, 1483–1489. [CrossRef]
10. Ciriello, G.; Miller, M.L.; Aksoy, B.A.; Sembabaoglu, Y.; Schultz, N.; Sander, C. Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. 2013, 45, 1127–1133. [CrossRef]
11. Amar, D.; Izraeli, S.; Shamir, R. Utilizing somatic mutation data from numerous studies for cancer research: Proof of concept and applications. Oncogene 2017, 36, 33–75. [CrossRef] [PubMed]
12. Yuan, Y.; Shi, Y.; Li, C.; Kim, J.; Cai, W.; Han, Z.; Feng, D.D. DeepGene: An advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinform. 2016, 17, 476. [CrossRef] [PubMed]
13. Ding, J.; Bashashati, A.; Roth, A.; Oloumi, A.; Tse, K.; Zeng, T.; Haffari, G.; Hirst, M.; Marra, M.A.; Condon, A.; et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. Bioinformatics 2012, 28, 167–175. [CrossRef] [PubMed]
14. Cai, Z.; Xu, L.; Shi, Y.; Salavatipour, M.; Lin, R.G. Using Gene Clustering to Identify Discriminatory Genes with Higher Classification Accuracy. IEEE Symp. Bioinform. BioEng. 2006, 6, 235–242.
15. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989, 2, 359–366. [CrossRef]
16. Cho, J.H.; Lee, D.; Park, J.H.; Lee, I.B. New gene selection method for classification of cancer subtypes considering within-class variation. FEBS Lett. 2003, 551, 5–7. [CrossRef] [PubMed]
17. Chen, Y.; Sun, J.; Huang, L.-C.; Xu, H.; Zhao, Z. Classification of Cancer Primary Sites Using Machine Learning and Somatic Mutations. BioMed Res. Int. 2015, 2015, 491–502. [CrossRef]
18. 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.; et al. TumorTracer: A method to identify the tissue of origin from the somatic mutations of a tumor specimen. BMC Med. Genom. 2015, 8, 58. [CrossRef]
19. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv* 2016, arXiv:1603.02754.
20. Katarzyna, T.; Czerwiska, P.; Wisnerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* 2015, 19, 68–83.
21. International Cancer Genome Consortium. International network of cancer genome projects. *Nature* 2010, 464, 993–998. [CrossRef] [PubMed]
22. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 1996, 58, 267–288. [CrossRef] [PubMed]
23. Sun, X. The Lasso and Its Implementation for Neural Networks. Ph.D. Thesis, National Library of Canada—Bibliothèque Nationale du Canada, Ottawa, ON, Canada, 1999.
24. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 2005, 67, 301–320. [CrossRef]
25. Yuan, M.; Lin, Y. Model Selection and Estimation in Regression with Grouped Variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 2006, 68, 49–67. [CrossRef]
26. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* 2013, 22, 231–245. [CrossRef]
27. Feng, J.; Noah, S. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv* 2017, arXiv:1711.07592.
28. Chollet, F. Keras, Online. 2015. Available online: https://github.com/fchollet/keras (accessed on 1 April 2019).
29. Abadi, M.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, Online. 2015. Available online: http://tensorflow.org (accessed on 1 April 2019).
30. Diederik, P.K.; Jimmy, B. ADAM: A method for stochastic optimization. ICLR. *arXiv* 2014, arXiv:1412.6980.
31. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 2002, 16, 321–357. [CrossRef]
32. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In ICIC Advances in Intelligent Computing; Springer: Berlin, Germany, 2005; pp. 878–887.
33. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the IEEE International Joint Conference on Neural Networks, Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
34. Tomek, I. Two Modifications of CNN. *IEEE Trans. Syst. Man Commun.* 1976, 6, 769–772.
35. Hart, P. The condensed nearest neighbor rule. *IEEE Trans. Inf. Theory* 1968, 14, 515–516. [CrossRef]
36. Kubat, M.; Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In Proceedings of the 14th International Conference on Machine Learning, ICML, Nashville, TN, USA, 8–12 July 1997; pp. 179–186.
37. Wilson, D.I. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Trans. Syst. Man Cybern.* 1972, 3, 408–421. [CrossRef]
38. Levina, E.; Bickel, P.J. Maximum likelihood estimation of intrinsic dimension. *Proc. NIPS* 2004, 1, 777–784.
39. Vogelstein, B.; Papadopoulos, N.; Velculescu, V.E.; Zhou, S.; Diaz, L.A.; Kinzler, K.W. Cancer genome landscapes. *Science* 2013, 339, 1546–1558. [CrossRef] [PubMed]
40. Hofree, M.; Shen, J.P.; Carter, H.; Gross, A.; Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* 2013, 10, 1108–1115. [CrossRef] [PubMed]
41. Kim, Y.M.; Poline, J.B.; Dumas, G. Experimenting with reproducibility: A case study of robustness in bioinformatics. *Gigascience* 2018, 7, giy077. [CrossRef]
42. LeMorvan, M.; Zinovyev, A.; Vert, J.P. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Comput. Biol.* 2017, 13, e1005573. [CrossRef]
43. Auslander, N.; Wolf, Y.I.; Koonin, E.V. In silico learning of tumor evolution through mutational time series. *Proc. Natl. Acad. Sci. USA* 2019, 116, 9501–9510. [CrossRef]