IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 March 2023, accepted 4 April 2023, date of publication 13 April 2023, date of current version 19 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3266983

RESEARCH ARTICLE

DrOGA: An Artificial Intelligence Solution for Driver-Status Prediction of Genomics Mutations in Precision Cancer Medicine

MATTEO BASTICO[1], (Student Member, IEEE), ANAIDA FERNÁNDEZ-GARCÍA[1], (Student Member, IEEE), ALBERTO BELMONTE-HERNÁNDEZ[1], (Member, IEEE), AND SILVIA URIBE MAYORAL[2]

[1]Escuela Técnica Superior de Ingenieros de Telecomunicacións, Universidad Politécnica de Madrid, 28040 Madrid, Spain
[2]Escuela Técnica Superior de Ingeniería de Sistemas Informáticos (ETSISI), Universidad Politécnica de Madrid, 28031 Madrid, Spain

Corresponding author: Matteo Bastico (mab@gatv.ssr.upm.es)

This work was supported by the H2020 European Project Genomed4ALL (https://genomed4all.eu/) under Grant 101017549.

ABSTRACT Precision cancer medicine suggests that better cancer treatments would be possible guiding therapies by tumor's genomics alterations. This hypothesis boosted exome sequencing studies, collection of cancer variants databases and developing of statistical and Machine Learning-driven methods for alterations' analysis. In order to extract relevant information from huge exome sequencing data, accurate methods to distinguish driver and neutral or passengers mutations are vital. Nevertheless, traditional variant classification methods have often low precision in favour of higher recall. Here, we propose several traditional Machine Learning and new Deep Learning techniques to finely classify driver somatic non-synonymous mutations based on a 70-features annotation, derived from medical and statistical tools. We collected and annotated a complete database containing driver and neutral alterations from various public data sources. Our framework, called Driver-Oriented Genomics Analysis (DrOGA), presents the best performances compared to individual and other ensemble methods on our data. Explainable Artificial Intelligence is used to provide visual and clinical explanation of the results, with a particular focus on the most relevant annotations. This analysis and the proposed tool, along with the collected database and the feature engineering pipeline suggested, can help the study of genomics alterations in human cancers allowing precision oncology targeted therapies based on personal data from next-generation sequencing.

INDEX TERMS Genomics, mutation, artificial intelligence, machine learning, deep learning, explainable AI, driver-status prediction, oncology, precision cancer medicine.

I. INTRODUCTION

Current efforts in precision oncology largely focus on the benefit of genomics-guided therapy. Multiple types of data are used to gather patients into groups that will most likely respond to a given treatment [1], [2]. Precision medicine fully relies on biomarkers in order to classify the patients risk status, its prognosis and the response to treatments [3]. Biomarkers are objective indications of the medical state observed outside the patient that can be measured accurately and reproducibly [4]. Among all, 'omics' data, such as DNA, RNA or proteins, allow the identification of putative biomarkers which are validated determining the range of conditions under which they will provide reproducible and accurate data [5], [6]. A lot of effort has been put into the production of affordable technology for '-omics' data retrieval [7], [8] but the validation of biomarkers is still stymied by low statistical power and poor reproducibility of results [9].

Thanks to the current advances in DNA sequencing techniques [10], [11], the amount of data available for research purposes exponentially increased in recent years. Next-generation sequencing (NGS) techniques have rapidly evolved over the past 15 years and new methods are continually being commercialized to allow high-throughput DNA

The associate editor coordinating the review of this manuscript and approving it for publication was Kin Fong Lei.

sequencing [12]. However, a standard pipeline for processing and analyzing this large amount of data has not been defined yet, arising inconsistencies and debates [13]. The best practices for the read-to-variant discovery workflows for germline and somatic short genomic variants are presented in [14]. In recent years, sequenced DNA reads are generally aligned to the reference genome, i.e., GRCh37 (hg19) or GRCh38 (hg38), though a *seed-and-extend* approach [15], such as bwa-mem [16] and bowtie2 [17]. Post-processing is used on aligned reads to remove PCR duplicates [18] and to perform Base Quality Score Recalibration (BQSR) [19].

Genomics-guided personalised cancer therapies require the identification of biomarkers, such as recurrent point mutations, translocations and potentially new therapeutic targets [20]. In this scenario, the discovery of somatic and germline Single Nucleotide Variants (SNV), as well as insertion-deletion mutations (indels), is vital. To this purpose, several algorithms, applied to aligned DNA reads, have been developed using probabilistic, mathematics, and Artificial Intelligence (AI) techniques [21], [22]. Based on several probabilistic models for genotyping and filtering, Mutect2 has been proposed to detect SNVs and indels [23]. Furthermore, deep convolutional neural networks can also call genetic variation in aligned NGS read data, as shown in DeepVariant [24].

Despite the flexibility and availability of variant calling techniques, the identification of mutation driver status as biomarker information remains challenging, especially in Machine Learning (ML) [25]. Nowadays, the frequency of a mutation in patients is still one of the most reliable indicators of the mutation driver status [26]. Several metrics and scores have been defined in recent years to quantify the pathogenicity of somatic mutations [27]. As shown in Figure 1a and Figure 1b the individual driver status classification tools have weak performance in terms of precision (less than 60%), in favor of higher recall, when applied to a generic dataset of somatic SNV, like ours. From a clinical perspective, an improvement in the precision of such algorithms would lead to more accurate driver status classifiers, reducing the workload of clinicians and doctors analysing manually all the patient-specific genomics alterations [28].

In this work, we first collect a dataset of driver status-annotated somatic non-synonymous variants joining various public sources. We propose a new features engineering pipeline in order to represent each mutation with a 70-features vector, obtained form several functional annotations, which is particularly suited for subsequent classical ML and Deep Learning (DL) algorithms. A new traditional ML and DL framework, called Driver-Oriented Genomics Analysis (DrOGA), is then proposed, aiming to improve the precision of the driver status prediction while keeping the recall unaltered. We show that our method gives the best performance on our data compared to individual and other ensemble methods. Finally, the results are explained with the help of eXplainable Artificial Intelligence (XAI) techniques in order to give a clinical perspective and model understanding. In clinical applications, the repercussion of the decision-making process can be critical for a patient [29]. XAI techniques are therefore introduced in this work because a deeper understanding of the models behaviour, usually seen as ‘‘black boxes’’, is the key to detecting the causes of failures and improving their performance [30].

II. RELATED WORKS

One of the most up-to-date tools for functionally annotating genomic alterations is ANNOVAR [31], an open software providing more than 100 features from gene, region and filter-based annotations, among others. These annotations are mostly retrieved from well-known genomics databases, such as dbSNFP [32].

Nevertheless, there are studies focused on creating benchmarks to prioritize variants annotations, matching some of the ones also provided by ANNOVAR [33], [34], [35]. Their main objective is usually to identify missense somatic variants that may be relevant as disease cause. These researches are usually framed within a defined area of study in which the tools are tested on specific datasets with a target disease. Neurodegenerative diseases are targeted in [33] using 18 different annotations, the majority of which are included, in an updated version, in our work. In [34], 24 tools are instead employed to identify the most sensitive ones according to the phenotype of the disease. Recently, a comparison of the performance of several annotation tools over a manually curated oncogenic dataset is performed by Suybeng et al. [35]. Here, as in our work, dbSNFP is one of the main annotations databases involved, and the final purpose of the study is differentiating between oncogenic and neutral mutations. It presents a comparison of single annotations and combinations of them taken two by two, without dealing with the contribution of the entire set of features or trying to develop a model to achieve the desired prediction from them.

Other publications go a step further along this line of research by creating their own prediction tools. One of the first annotation tools specialised in exploring functional effects in cancer somatic mutations is Functional Annotation of Somatic Mutations in Cancer (FASMIC) [36], based on cell viability and Reverse-Phase Protein Arrays (RPPAs) assays. In 2019, Variant Call Format-Diagnostic Annotation and Reporting Tool (VCF-DART) [37] was released as a method to identify genetic variants that may be of clinical importance based on a custom gene list and annotation tools such as dbNSFP, SNPSift [38] and Variant Effect Predictor (VEP) [39].

With the objective of creating prediction tools to identify driver cancer mutations, traditional ML algorithms have also been proposed in other works. Cancer-specific Driver missense mutation Annotation (CanDrA) [40] was already proposed in 2013 and trained with samples mainly obtained from The Cancer Genome Atlas (TCGA) [41] and the Catalogue Of Somatic Mutations In Cancer (COSMIC) [42]. A weighted Support Vector Machine (SVM) classifier is trained with 95 features per sample from annotation portals such as VEP

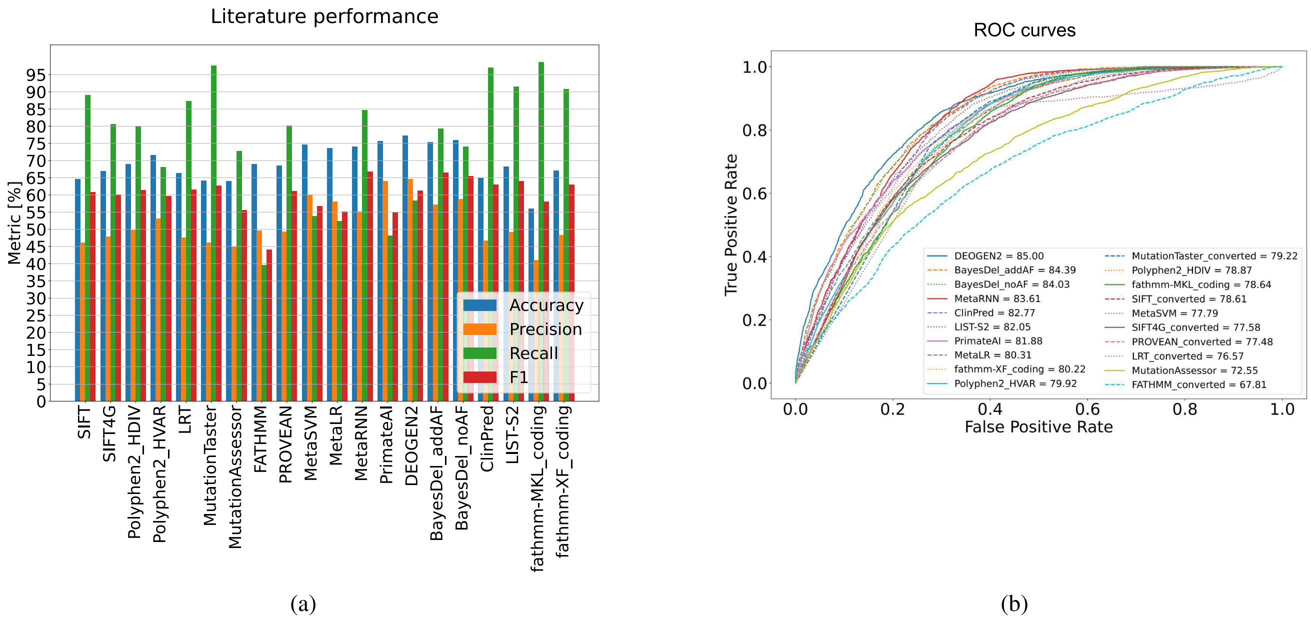


FIGURE 1. Statistics of State-of-The-Art conservation, functional and ensemble methods for driver status prediction on our test set. All of them are used in a rankscore scaled version as input of our pre-processing pipeline and classification benchmark. (a) Accuracy, Precision, Recall and F1 score of those algorithms providing prediction labels ordered from left to right according to F1. Most of them reach good performances in recall but precision and accuracy are low. (b) ROC curves and AUC of the same algorithms based on the deleterious score ranging between 0 and 1, where an high value corresponds to high probability of oncogenicity.

and ANNOVAR. An SVM architecture is also employed in Combined Annotation-Dependent Depletion (CADD) [43] based on a dataset of more than 20 millions of variants annotated with 62 features each. Its architecture has been updated more recently to a Deep Neural Network (DNN), improving performance with the so-called Deleterious Annotation of genetic variants using Neural Networks (DANN) [44]. BayesDel [45] implements a naïve Bayesian model within an extended gene-prioritisation framework. A large variety of approaches proposed in the literature are tree-based models that gather other ensemble and functional predictions. Among all, DEOGEN2 [46] uses a Random Forest (RF) architecture, ClinPred [47] implements RF and XGBoost (XGB), and CHASMplus [48] a RF tool specialised in predicting driver mutations in 32 cancer types. Kernel-based models are instead proposed in FATHMM-XF [49] and Recurrent Neural Networks (RNN) in MetaRNN [50]. However, all of the presented methods have good recall in the driver status prediction in our data set, but low precision, as shown in Figure 1.

Lastly, AIDriver [51] is the most similar research to our proposal. Its aim is to predict the driver status of somatic missense mutations based on 23 pathogenic phred-scaled features, all of them included in dbSNFP. They compare several classical ML model architectures like XGB, SVM, RF, and Multi-Layer Perceptron (MLP), obtaining the best performance with XGB. According to their results, this method outperforms the latest versions of previous algorithms, using test data from different sources like TGCA, Cancer Genome Interpreter (CGI) [52], and International Cancer Genome Consortium (ICGC) data portal [53].

Our proposal is intended to improve the driver status prediction performance by taking advantage of all the available information retrieved by ANNOVAR. Most of the features used in AI-Driver are also considered in this work with a more recent normalization technique performed over the complete dbSNFP dataset. Additional annotations are also gathered from dbSNFP and several other databases. Moreover, we provide a general data cleaning and preparation pipeline for annotations collected from these databases, and a complete referenced clinical documentation. Several traditional ML architectures are also tested on our features representation, including new DL approaches such as Convolutional Neural Networks (CNN). Furthermore, merging clinical information with XAI methods is the key to understanding the underlying behavior of our models. This issue is often barely covered by previous tools due to the poor self-explaining power of ML and DL techniques.

III. MATERIALS AND METHODS

We generated our training and testing dataset gathering driver and neutral or passengers somatic non-synonymous, e.g. missense, mutations publicly provided by various data sources. Namely, we collected 2477 driver mutations from CGI [52], 355 driver and 460 passengers mutations from FASMIC [36], [56]. In the latter, as specified by the authors, we considered as driver mutations the variants labeled as activating, inactivating, inhibitory, non-inhibitory, while, as passengers, the ones labeled as neutral. Subsequently, single nucleotide polymorphism (SNP) identifiers of all the somatic missense variants with Minor Allele Frequency (MAF) higher than 5% [58] were collected from dbSNP [57]. Firstly, we annotated them

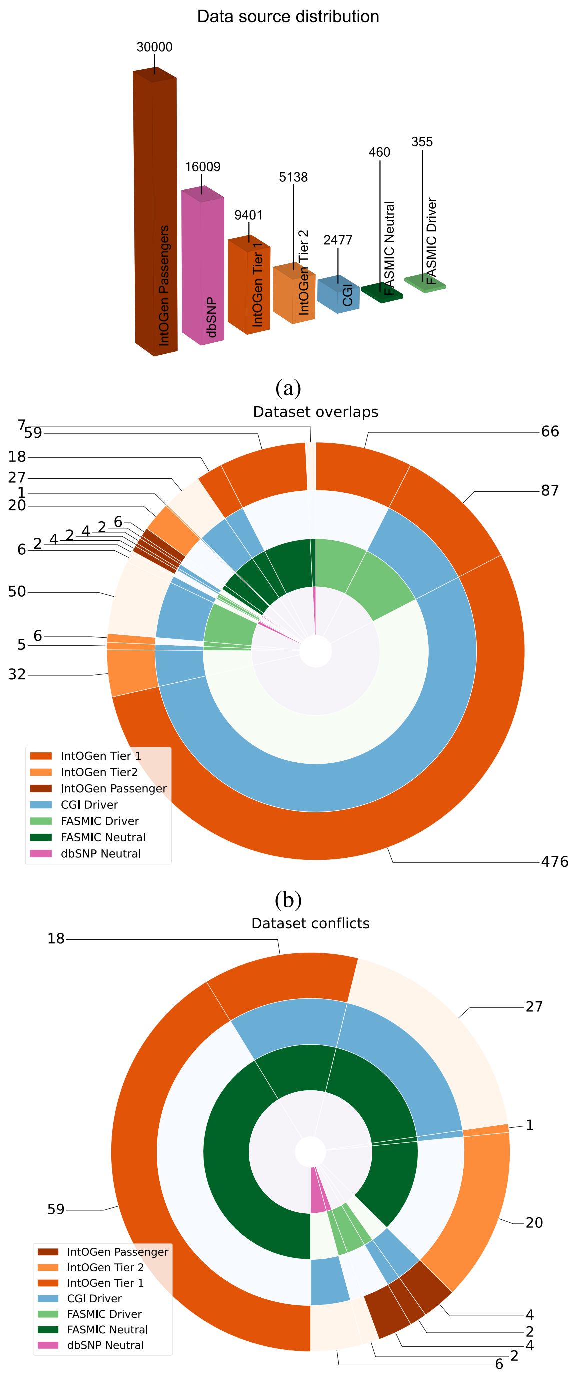


FIGURE 2. **Statistics of the collected dataset. (a) Number of mutations per data source, divided by category, gathered in our dataset.(b) Overlaps between the joined data sources: IntOGen [54], [55], CGI [52], FASMIC [36], [56] and dbSNP [57], from the outer ring to the inner one. (c) Focus on the conflicting overlaps. Some variants variants are annotated discordantly by the different data sources. The ground-truth values kept in our dataset are the most recent annotations (from inner to outer ring).**

with VEP [39] and we split multiallelic mutations in multiple biallelics [59]. Second, mutations were further annotated with ANNOVAR gene-based annotations and gnomAD v2.1.1 Alleles Frequencies (AFs) [60]. Only exonic or splicing non-synonymous mutations with gnomAD AF higher than 5% are kept as neutral mutations in our dataset. The double filtering is done to further ensure the neutrality of the collected variants, which are in total 16009. Finally, all the exonic or splicing non-synonymous mutations in IntOGen [54], [55] are also collected. Here, variants are labelled as Tier 1, Tier 2, passengers and not protein-affecting. Tier 1 and Tier 2 are mutations with strong clinical significance and potential clinical significance, respectively [61]. For that reason, we consider both as positive cases, i.e. drivers, while we use the passengers variants as negative cases. In total, we collected from IntOGen 9401, 5138 and 30000 Tier 1, Tier 2 and passenger mutations, respectively. Neutral variants are randomly sampled among a total of 454722 variants to avoid extreme unbalancing on the dataset. The number of collected variants per data source divided by category are summed up in Figure 2a.

Identical mutations among the different sources were found in the gathered data, as shown in Figure 2b. Although most of the overlapping samples match ground truth and can be dropped without further actions, there are more than 100 mutations presenting a different classification depending on the source from which they are collected (focus in Figure 2c). In order to not dispense with them, the label retrieved from the most recent version database is kept as true label, giving preference to the samples from dbSNP, as they have been already filtered twice. Therefore, the priority is led by dbSNP, followed by FASMIC (2021), CGI (2018), and IntOGen (2016), i.e., from inside to outside of the Figure 2c representation.

Finally, the merged dataset is set up as VCF to be annotated by ANNOVAR. The annotations are retrieved from the National Center for Biotechnology Information (NCBI) [62], University of California Santa Cruz (UCSC), Genome Browser [63], gnomAD, Exome Aggregation Consortium (ExAC) [64], dbSNP, dbSNFP v4.2a [32], [65], ClinVar [66], COSMIC, American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) [67], ICGC, Kaviar [68], Haplotype Reference Consortium (HRC) [69], the 1000 Genomes Project Consortium [70] and NHLBI Exome Sequencing Project (ESP) [71], together with some further information belonging to ANNOVAR. Every feature from those annotation databases is deeply analysed and classified to later perform a proper and meaningful feature selection as input for our model building. The detailed study of each annotation can be found in the documentation of the dataset, including a general description of each field, its category, possible values or range, additional information, and original source.

Overall, the proposed dataset is a 196-features collection, including 7 mandatory VCF fields and 189 external annotations, gathering 16360 driver samples and 46444 neutral or passenger mutations. The training and testing split is performed by randomly sampling with an 80%-20% proportion the original dataset, resulting in 50243 training samples, of which 13086 are driver, and 12561 for testing, containing 3274 positive cases.

Based on the collected dataset, we provide and propose the following contributions for the research community.

- ANNOVAR Features Explanation and Details: We include a complete documentation of the dataset where each feature has been analysed and commented rigorously providing a simple explanation, the type of data and its possible values with some links to resources for a deeper understanding.
- Dataset and code for Feature Extraction and Modelling: Along the paper, we provide all the code and materials to help the research community to preprocess the data following the proposed pipeline and to replicate the final results of our driver status prediction benchmark (see Section VII).
- Comparison of Different Models for Driver Status Prediction: A benchmark with different classical ML and DL models is proposed to compare the performance of different methods in the driver status prediction of somatic non-synonyms mutations.
- Deeper Analysis Based on XAI: In order to understand the results and try to link them with a real interpretation from a clinical perspective, we analyse our models with XAI to understand their underlying behaviors.

IV. GENERAL APPROACH

In this section, we provide details on the proposed techniques to filter, process and analyse our annotated dataset of somatic non-synonymous mutations. Firstly, our feature engineering and data cleansing pipeline to preprocess the data is deeply explained. After that, various models for driver status classification are introduced. Finally, the used techniques to get explainable insights on the proposed traditional ML and DL models are presented with the aim of obtaining an interpretation in a clinical perspective.

A. FEATURE ENGINEERING AND DATA CLEANSING

Originally, 189 features per mutation were annotated in our database as output of ANNOVAR. The annotations are classified in three groups according to the information they provide: (1) Allele Frequency (AF) in a specific database, (2) filter-based pathogenic predictors that can be divided, according to the nature of their data, as categorical, numeric or binary statistics, and (3) other kinds of knowledge, such as external database identifiers or gene-based information.

1) AF ANNOTATIONS

Raw frequencies values for each variant are retrieved from gnomAD, ExAC, Kaviar, HRC, 1000 Genomes and NHLBI ESP databases. Generally, the raw frequencies of alleles at a particular locus are only available for common variants in a population, such as SNPs. Moreover, AF information is already included, with further processing, in some filter-based annotations, e.g., MVP [72], ClinPred [47] and PrimateAI [73]. Therefore, to avoid information redundancy and missing input data for rare mutations, we remove all the raw AF from our features vectors. This category presents a total of 37 dropped annotations.

2) FILTER-BASED PATHOGENIC PREDICTORS

In order to simplify the data preparation and avoid parsing steps, numerical pathogenic indicators are prioritized over categorical ones belonging to the same research. In most of the features provided by dbSNFP, which encompasses more than half of our original annotations, the rankscore annotation has been included since 2014 [74]. It is a normalised score among the whole source dataset, ranging between 0 and 1. These annotations are also preferred over categorical and not-normalised ranges items, as they ease the data cleansing process. Within this category, there are 39 rankscore variables considered as final input features, replacing a total of 61 non-normalised annotations providing the same information. Namely, the kept filter-based pathogenic predictors rankscores can be classified by their nature as follows:

- Conservation scores: GERP++ RS [75], LRT [76], PhyloP (100 and 30way) [77], PhastCons (100 and 30 way) [78] and SiPhy [79].
- Functional prediction scores: SIFT, SIFT4G [80], MutationTaster [81], MutationAssessor [82], [83], FATHMM [84], PROVEAN [85], [86], VEST4 [87], LIST-S2 [88], MVP [72], MPC [89], integrated fitCons [90], PrimateAI [73], PloyPhen2 (HDIV and HVAR) [91] and MutPred [92].
- Ensemble scores, with the objective of predicting deleteriousness, based on collections of features already available in dbSNFP: MetaSVM, MetaLR [93], MetaRNN [50], BayesDel (with and without AF) [45], ClinPred [47], REVEL [94] and M-CAP [95], together with other ensemble scores of features non-belonging to this database: DEOGEN2 [46], CADD [43], DANN [44], FATHMM-MKL [96], FATHMM-XF [49], Eigen (raw and PC) [97], GenoCanyon [98] and LINSIGHT [99].

Nevertheless, there are categorical and numerical annotations not transformed nor normalised which are not associated to a rankscore. The treatment suggested by this pipeline for the numerical values is applying min-max normalization within the defined range or, if it is not predefined, standardising the data. This is applied for two features from the dbSNFP database: confidence value of fitCons [90] and GERP++ NR [75]. On the other hand, the categorical features are parsed to a single integer value if multiple-class for a single sample is not allowed. On the contrary, when there are possible multiple categories for one sample, one-hot encoding [100] is implemented. This process has been carried out on the two dbSNFP variable, i.e. ALoFT [101] and another pathogenic feature from ClinVar.

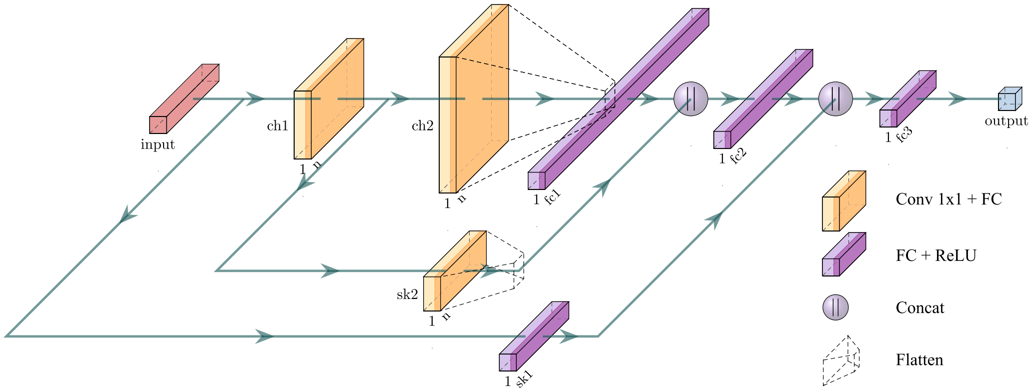


FIGURE 3. Convolutional network architecture for genomics variant classification with skip connections. The overall schema is an encoder-decoder architecture. The input is fed simultaneously into a 1 × 1 convolution and a fully connected layer used as first skip connection. The second one is inserted after the first convolutional layer. Their output features are used to transfer additional information in the decoding stage which is composed of fully connected layers. The normal convolutional architecture is obtained removing the skip connections.

3) OTHER KIND OF KNOWLEDGE

Regarding the external database identifiers and gene-based information, a simplification of the annotations is kept. There are several identifiers from external resources like Genotype-Tissue Expression (GTEx) Program [102], InterPro database [103], ICGC, COSMIC and ClinVar. Each feature is transformed into a binary value, being positive in the cases in which such mutation is registered in the given database. 5 features are obtained from them, while additional information from the same databases and gene-based studies is deleted in order not to bias the driver status prediction on the gene-based information, i.e., chromosome or location. This comprises 8 dropped annotations. Binary annotations without a related rankscore are originally from ACMG/AMP, and they are kept as retrived, because no further transformation is needed. 28 features are included here, and one containing duplicate information is dropped. Finally, the beginning and ending flags are deleted, together with 2 functional annotations, because non-synonymous exonic or splicing filters have already been applied in the dataset collection phase. VCF features are also dropped, together with the last single feature from USCS, as they do not provide pathogenic-related information and our data may have undesired dependencies and unbalances on location attributes.

Overall, 77 features are kept from the original annotations. A final cleansing is performed after splitting the one-hot encoding variables into single columns. Every column not providing any information for our training data, i.e. assigning the exact same value for every sample, is dropped. These are a total of 17 features to be discarded. Additional columns which present an high probability of missing annotations, are also disregarded (LINSIGHT, M-CAP, MutPred and MVP). Moreover, samples with missing annotations are filtered out. The guideline we propose is to eliminate samples with at least one rankscore or one binary ClinVar feature missing, as they are the main pathogenic attributes. The remaining null annotations have been filled with 0 or -1 values, for convenience. In [51], the missing input features are estimated by averaging the values of near-variants or of the whole database. The rankscores and the ClinVar predictions have an important clinical significance which often differs from near variants. For this reason, we decided to discard examples with missing annotations instead of estimating them. Ultimately, there are 32574 training samples, with 9993 positive cases, and 8102 test samples, containing 2500 driver labels, each sample consisting of a 70-feature vector based on ANNOVAR annotations. As Supplementary Data, we include the code to perform the data preprocessing and cleansing proposed here.

B. PROPOSED MODELS FOR DRIVER-STATUS PREDICTION

The main objective of our benchmark is to classify between variants detected as driver for cancer evolution and non-driver mutations, based on the collected ANNOVAR features pre-processed as we proposed in Section IV-A.

Several AI models have been introduced in the literature with supervised classification purposes [104]. Depending on the input data format, different techniques can be suitable to properly solve the task. In this work, we propose the use of not only classical ML architectures but also DL architectures to perform an exhaustive analysis on both types of algorithms. The list of models tested in our benchmark is the following:

- Traditional Machine Learning (ML): Logistic regression (LR), SVM, Decision Tree Classifier (DTC), RF and XGB.
- Deep Learning (DL): Deep MLP and Convolutional Neural Network (CNN).

The newly proposed CNN architecture is shown in Figure 3. The rationale behind is to use an inverse encoder-decoder schema such that the input features vector is represented in an higher dimensional space which is then used to obtain the final classification. 1 × 1 convolutional layers [105], [106] are used to increase the filter space dimensionality, i.e. to transform each input feature independently.

The features higher dimensional representations are then flattered and fed into fully connected layers in order to retrieve the driver-status prediction. Modified skip connections [107] are introduced to pass compressed lower-level semantic information to the latter layers of the neural network. Their aim is to keep more original information regarding the input features vector in the decoding phase. The CNN model is trained and tested with and without skip connections, in order to evaluate possible performance improvements.

Traditional ML and DL algorithms require the tuning of several non-learnable parameters, e.g. learning rate, in order to choose the optimal configuration for the learning process. Model hyper-parameters optimisation [108], [109], [110] is performed during training in order to select the best configurations for the proposed solutions. The hyperparameter optimization technique that we used for classical ML models in our benchmark is called Random Search [111]. Random configurations of the model hyper-parameters are tested within a range of selected values and the optimisation is done on the resulting F1-value. Regarding DL models, hyperparameters tuning is done using the Adaptive eXperimentation (AX) algorithm, based on Bayesian optimisation, and Asynchronous Successive Halving (ASHA) scheduler [112] is included to improve the search efficiency. Due to the unbalancing of the dataset, we decided to include the loss as hyper-parameter of our DL models. Separately, we tested the weighted Binary Cross Entropy (BCE) Loss and the α-balanced variant of the Focal Loss (FL) for binary classification [113]:

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{1}$$

where

$$p_t = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{otherwise} \end{cases} \tag{2}$$

p is the predicted output and y the ground-truth. By tuning the parameters γ and α_t, it allows to focus on hard examples and to balance the loss equally for both classes, respectively. The configuration of the search spaces and the resulting best models for both traditional ML and DL are reported in the GitHub repository along with the code. To evaluate the driver status prediction performances of each model, we provide several well-known metrics such as Accuracy, Precision, Recall, F1-score and Confusion Matrix [114].

C. INTERPRETABILITY WITH XAI

In medicine and especially in genomics, model interpretability is a key point to match AI with medical interpretation of the annotations under study used as input features of the algorithms [115]. In this work, our models are analysed with XAI techniques to understand their hidden behaviour and the Feature Importance (FI).

In some linear models, the FI can be directly inferred from the parameters learned during training. Their values are used to analyse each feature separately and to quantify its contribution when a new prediction is made. In our benchmark, the models that provide this information are LR and SVM, when a linear kernel is used. DTC also provides directly the FI based on the number of splits created in the training phase. In this way, it is possible to reproduce the entire tree and understand the model decision.

In other cases, local surrogate models are interpretable models that are used to explain the individual predictions of "black-box" models [116], [117]. Among many of the techniques available for XAI, SHapley Additive exPlanations (SHAP) [118] has been selected in this work due to its versatile analysis on feature relevance, such as global or local importance. In addition, this technique is suitable for feature interaction inspection and can be used with complex model architectures such as neural networks. It is an implementation that connects optimal credit allocation with local explan ations using the Shapley values from game theory. It applies to all the models of our benchmark, such as XGB or CNN, considering the input as a features vector or an image, respectively. Therefore, our interpretation of the models and their connection to clinical significance is done through SHAP analysis.

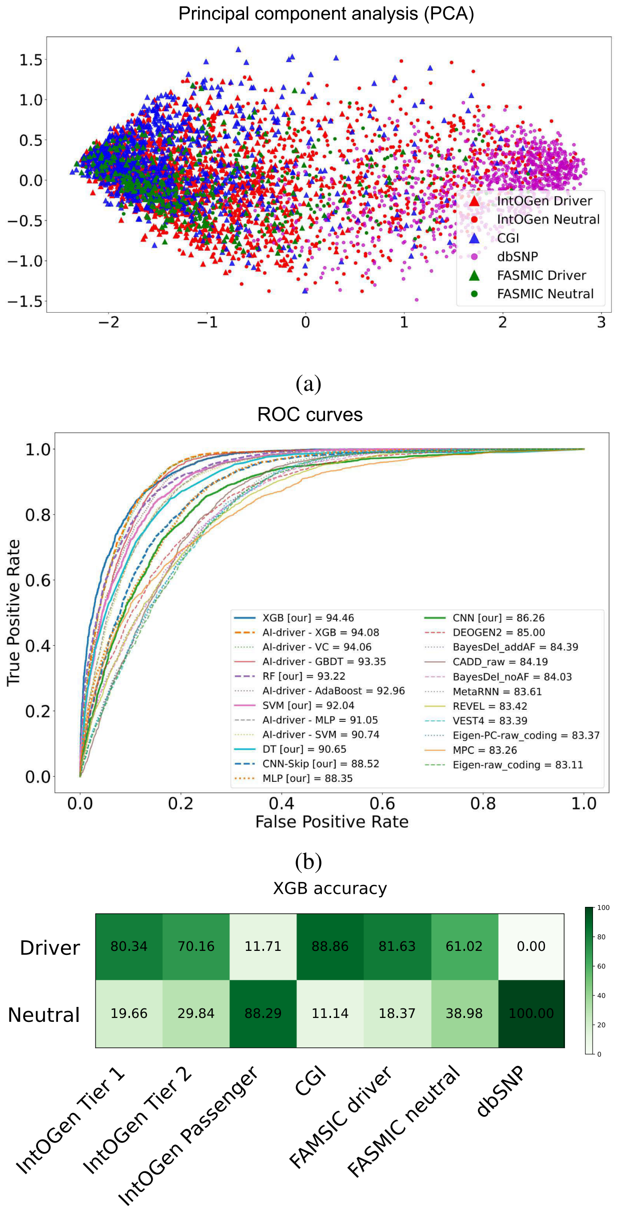
V. RESULTS AND DISCUSSION

This section is divided into two main parts. First, the results obtained with our feature engineering and benchmark, and their comparison with other works are presented. Second, an extensive research over some selected algorithms is included to analyse the Feature Importance (FI), based on XAI.

A. MODELS TRAINING AND EVALUATION

As preliminary result of our features engineering pipeline, and to test in advance the power of our final 70-features representation of DNA variants, we applied a Principal Component Analysis (PCA) preserving only two dimensions. The distribution in the 2D-space of a random subset of 3815 samples of our training samples, containing 1000 from each database and all samples from FASMIC, is shown in Figure 4b. Driver and neutral mutations tend to be placed on opposite sides of the space, giving a somehow clear distinction between them. This insight shows that our variant codification is effective in being used as input for further analysis, as shown in the following.

Table 1 collects the results of our benchmark for driver status prediction of somatic non-synonymous mutations in terms of F1-score, Recall, Precision and Accuracy. The scores are compared with State-of-the-Art algorithms both in DL, classical ML, and clinical fields. In order to provide a fair comparison, the AI-driver algorithms [51] are re-trained and tested using our clean raw dataset. The approach proposed by that solution is applied over the dataset for training their models, obtaining a 23 feature-vector. Our models outperform in terms of F1-score, giving the best balance between precision and recall. The inclusion of more annotations as input improves the performances, allowing to achieve, in most of the cases, better accuracy, precision, and recall.



(a)

(b)

(c)

FIGURE 4. (a) Two components PCA analysis of our 70-feature representation for DNA variants. The effectiveness of the representation is clear since driver mutations (triangles) are mostly distributed on the left side, while neutral variants (circles) are on the right. (b) ROC curves and corresponding AUC of the proposed benchmark for variant driver status prediction. Other solutions are also reported as comparison and the best AUC is obtained with our XGB model. (c) Driver status predictions of the XGB model on our test set. The percentage of mutations classified as driver or neutral is shown for each collected data source, in order to be directly compared with the ground-truth.

The best F1-score is given by XGB which provides similar results both in precision and recall. The CNN with skip connections proposed in this work is instead able to increase the recall up to almost 90%, keeping a good trade-off with precision and accuracy scores. Classical solutions, such as FATHMM-XF [49], ClinPred [47] and LIST-S2 [88], can reach similar recall value at the expense of a big drop in precision, i.e. more than 20% less with respect to our model. Improvements in SVM and RF are also obtained using our set of features. Nevertheless, in SVM, more False Positives are detected, reducing its precision compared to other methods. RF, instead, achieves the second-best performance in the traditional ML category thanks to the ensemble of several trees where different combinations of features are considered as input.

Figure 4a shows the ROC curves and AUC of our benchmark algorithms compared to other variants driver status classification methods, both ensemble and functional scores. In terms of AUC, our XGB algorithm is the best performing solution, showing that the inclusion of more input features increases the True Positive Rate (TPR). This helps to better discriminate between driver and neutral mutations, giving more reliable outputs for clinical analysis. Nevertheless, DL models do not exceed the results obtained with classical ML models due to the reduced number of input samples available for training.

Finally, Figure 4c shows the percentage of variants classified as driver or neutral by the XGB algorithm for each data source in our test set. The classification accuracy is lower for IntOGen Tier 2 (70.16%) and FASMIC neutral (38.98%) mutations. In the first one, the alterations are associated with a potential clinical significance, without showing a clear pathogenicity, reflecting in weaker performances in our driver status prediction models. On the other hand, FASMIC neutral variants, giving a closer look at the PCA decomposition in Figure 4a, are distributed in the features space like driver mutation, affecting directly their prediction accuracy. It is important to notice that the latter only comprises for the 0.72% of the total data and, therefore, it does not affect the overall performances of our benchmark.

B. XAI FOR MODEL UNDERSTANDING

After the analysis of the evaluation metrics, a deeper analysis is needed in order to give a clinical interpretation of our models. The FI are of particular interest to understand which are the most relevant annotations when performing driver status predictions. Given that, it is possible to discover the features that are affecting positively or negatively the algorithms and the reasons why mutations are correctly or wrongly classified.

The SHAP values of the input features for XGB and our CNN with skip connections are shown in Figure 5. For the first, the mean absolute SHAP values and the raw SHAP values per sample, along with the feature magnitudes are depicted in Figure 5a and Figure 5b, respectively. Regarding CNN, the average SHAP values for True Positives (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs) are shown in Figure 5c, considering the input vector as a one dimensional image. A complete analysis of the FI for every model of our benchmark can be retrieved from our source code. Overall, the greater is the SHAP value for an input feature, the higher is the impact of the corresponding annotation in the final classification. On the other hand,

TABLE 1. Results comparison in terms of F1 score, Recall, Precision and Accuracy. The algorithms are divided into three subcategories: (1) Deep Learning and (2) Machine Learning (3) Clinical. For each category, the best performing algorithm is shown in bold, while the overall best is underlined.

		F1	Recall	Precision	Accuracy
DL	CNN-Skip [our]	**76.84**	**89.24**	67.46	83.40
	CNN [our]	76.36	87.60	67.68	83.26
	MLP [our]	75.64	85.64	**67.73**	82.98
	AI-Driver - MLP [51]	71.79	78.77	65.95	**83.92**
	MetaRNN [50]	65.05	83.97	53.09	75.71
Traditional ML	XGB [our]	**79.20**	78.52	**79.89**	87.27
	RF [our]	76.39	74.88	77.97	85.72
	AI-Driver - VC [51]	76.27	79.23	73.52	87.19
	SVM [our]	76.22	79.80	72.94	84.63
	AI-Driver - XGB [51]	76.12	78.06	74.28	**87.28**
	AI-Driver - RF [51]	75.50	78.43	72.77	86.78
	LR [our]	74.25	73.88	74.62	84.18
	AI-Driver - GBDT [51]	73.58	74.42	72.76	86.12
	AI-Driver - AdaBoost [51]	73.30	75.38	71.33	85.74
	DT [our]	72.88	71.48	74.33	83.58
	AI-Driver - SVM [51]	69.95	71.06	68.89	84.15
	BayesDel [45]	64.90	78.66	55.24	76.75
	FATHMM-XF [49]	61.08	90.30	46.15	65.59
	ClinPred [47]	60.52	**96.94**	44.00	65.51
Clinical	LIST-S2 [88]	**62.28**	**91.27**	**47.27**	**67.50**
	PolyPhen [91]	59.04	80.21	46.71	63.89
	SIFT [80]	57.18	89.30	42.05	61.73

a positive value indicates that a feature influences the model towards a driver outcome, while a negative value pushes it toward a neutral prediction.

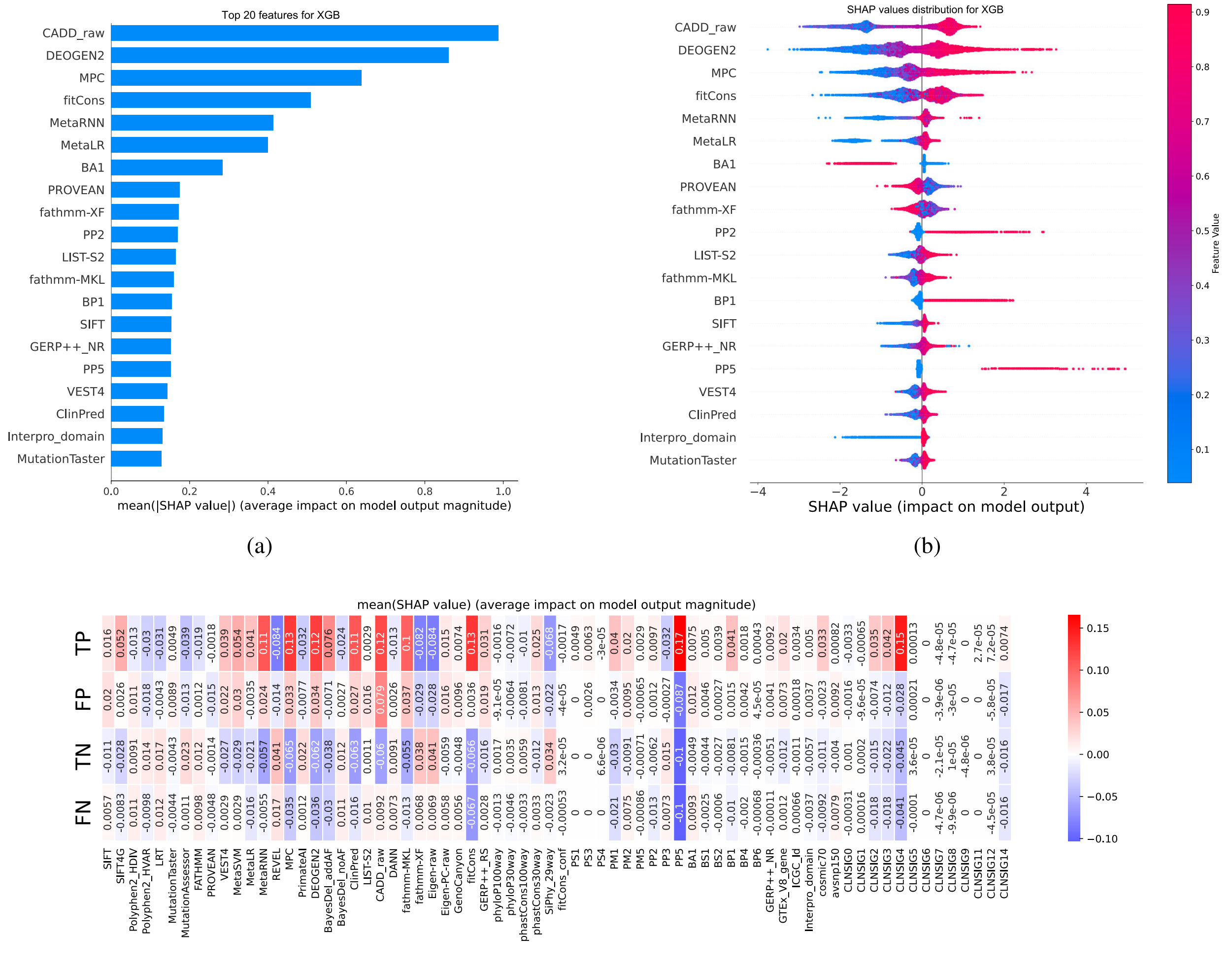
In general, there are specific annotations that stand out in FI over the rest of our 70-features input vector. DEOGEN2, CADD, MPC, and FitCons are always in the top-6 most important attributes. A part from FitCons, they were presented as some of the best unitary pathogenic predictors, according to the AUC in Figure 1b. Therefore, their qualities are learned by our models, which tend to give them more importance with respect to other annotations. On the other hand, the FI of other attributes, such as FitCons, is given by the fact that, following a preliminary study on our dataset, their distribution over the samples differs from the vast majority of input features. In this case, the models recognise them as peculiar annotations that deserve more importance in the final prediction. Additionally, both traditional ML and DL approaches contain within the 10 most important features several binary-ranged attributes such as PP5, PP3, BA1 and PP2. They are usually highly reliable, although hardly unbalanced annotations. Most of them tend to be disregarded by our benchmark due to its inconvenient distribution and the difficulty of the thresholding process in tree-based classifiers. The clearest example is given by the CLNSIG one-hot-encoded features, they barely participate in the decision-making process, but few of them may have high significance in some architectures, e.g. in CNN with skip connections of Figure 5c just 3 components over 13 have relevant SHAP values. Nevertheless, one of the main differences found between classical ML and DL models is the distribution of the FI, which is much more balanced among annotations in the latter category. Traditional ML architectures, mainly built up on tree-based classifiers, strongly route the predictions already in the first decisions steps, giving more importance only to the annotations involved in these phases. Instead, DL models have a wider view of the entire input features vector, being able to perform deeper combinations of the available attributes.

Focusing on Figure 5b, the general trend is that the rankscores of pathogenic predictors have a positive impact on the model output when their values are higher than 0.5. In contrast, benign predictors give negative SHAP values in correspondence of high scores. A deeper analysis of the classification errors can be carried out considering the SHAP values of TP, FP, TN and FN samples as separate groups (Figure 5c). This visualisation backs up the idea that TPs and TNs are well defined, but, usually, lower SHAP values are assigned to miss-classified samples. Moreover, FPs have lower but similar contributions to TPs, while FNs have some-how mixed contributions that do not clearly belong to them in a specific class. Nevertheless, there are some features, e.g. REVEL, FATHMM-XF, Eigen-raw and SiPhy, that seem to push the network towards a wrong decision for correctly classified examples, but at the same time they are useful to achieve better results on FPs and FNs. Such a detailed analysis can help tuning the annotations in order to improve performances on wrongly classified variants.

C. CLINICAL MEANING

Regarding the annotations providing an high FI, it is found that 2 of the top features, DEOGEN2 and CADD, are ensemble methods that collect information from different studies and databases with the aim of predicting the deleteriousness of variants [43], [46]. The fact that their sources are different makes them suitable to contribute together with valuable pathogenic information, without encompassing duplicated knowledge. MPC is also a driver predictor for missense variants which makes the difference with previous attributes by including frequency information, together with other deleteriousness predictors [89]. Similarly, fitCons is another functional annotation that includes fitness consequences of mutations based on conservational evaluations [90], providing an alternative perspective for classification.

Moving into the binary pathogenic predictors, BA1, PP2 and PP5 are all retrieved from ACMG/AMP [67]. PP2 and PP5 are described as supporting evidence of pathogenic indicators, based on pathogenic reports that include those variants and the gene rate containing benign or driver mutations, respectively. BA1, in contrast, is a benign classifier



(a)

(b)

(c)

FIGURE 5. SHAP values of the input features for XGB and CNN with skip connections of our benchmark. (a) The top 20 important features ranked by mean absolute SHAP values for the XGB model. (b) SHAP values per sample of the top 20 most important annotations for XGB model. The colour of each dot represents the original value of the input features. (c) Average SHAP values of TPs, FPs, TNs and FNs for our CNN with skip connections architecture, considering the feature vector as a one dimensional image.

that takes into account frequency information gathered from different databases, which is useful for well-known variants. Finally, the CLNSIG feature summarises different clinical significance approaches, including ACMG/AMP as well for pathogenic predictions [66]. That is why CLNSIG4, the 'pathogenic' class for the one-hot encoding of the original variable, provides information considering the same criteria.

VI. CONCLUSION

In this work we proposed a new features engineering pipeline to represent DNA variants by mean of a 70-feature vector collecting update-to-date conservation, functional and ensemble scores. The annotations are collected from several well-known databases, which can be easily accessed through the ANNOVAR annotating tool. With such processed data, we provide a benchmark, DrOGA, comprising classical ML and newly proposed DL architecture, to detect the driver-status of somatic non-synonymous mutations. In order to train and test our models, we collected a complete dataset of driver and neutral alterations from many public data sources.

We shown the effectiveness of our pre-processing pipeline and benchmark providing outperforming results in terms of variants driver-status classification, overcoming State-of-The-Art functional and ensemble methods. Ensemble methods generally surpass unsupervised methods when high-quality training data of appropriate type and quantity are available. This means that, based on our results, the proposed feature engineering helps to refine the quality of the data and obtain better predictions. Therefore, it can be used not only for the purpose of this work, but also to help other variant classification and prediction tasks, such as survival analysis or mutations clustering. On the other hand, there is a strong dependency related to the availability of the annotations used as input. The lack of some of them can compromise the correct functioning of the pipeline and subsequent algorithms. To overcome this problem, in future works, some complementary solution to estimate the missing values can be developed or included in the features engineering pipeline. For instance, the use of Variational AutoEncoders (VAEs) and Deep Latent Variable Models (DLVMs) has been already presented with the purpose of handling and imputing incomplete datasets in [119] and [120], respectively. Such approaches could be easily introduced as further pre-processing step in the proposed pipeline, to make possible the classification of every variant even when the set of annotations is not complete.

A final analysis through XAI is also provided to understand the hidden behavior of the models and to relate the outcomes with the clinical meaning of the annotations. We shown which features are more important for the final decision based on their SHAP values. This study can help future works to further refine the training and testing data removing or weighting variables which do not influence the models predictions. Overall, we believe that the proposed work can help the developing of tools to analyse genomics mutations in human cancers, allowing for further improvements in precision cancer medicine.

VII. SUPLEMENTARY DATA

The DNA mutations used in this work were gathered download from public repositories CGI [52], FASMIC [36], [56], dbSNP [57] and IntOGen [54], [55]. The merged data annotated with ANNOVAR, following our dataset collection, is available at https://github.com/matteo-bastico/DrOGA.

The code used to pre-process the collected dataset and obtain our variants representation is available in the same GitHub repository. An open source implementation in PyTorch and Scikit-learn of DrOGA is publicly accessible at the same repository, which is released under the MIT licence. It includes the independent training of all the presented architectures, as well as code for testing our pre-trained models on new variants.

REFERENCES

[1] D. Senft, M. D. M. Leiserson, E. Ruppin, and Z. A. Ronai, "Precision oncology: The road ahead," *Trends Mol. Med.*, vol. 23, no. 10, pp. 874–898, Oct. 2017.

[2] R. Nussinov, H. Jang, C.-J. Tsai, and F. Cheng, "Review: Precision medicine and driver mutations: Computational methods, functional assays and conformational principles for interpreting cancer drivers," *PLOS Comput. Biol.*, vol. 15, no. 3, Mar. 2019, Art. no. e1006658.

[3] A. Sarma, C. S. Calfee, and L. B. Ware, "Biomarkers and precision medicine: State of the art," *Crit. Care Clinics*, vol. 36, no. 1, pp. 155–165, 2020.

[4] K. Strimbu and J. A. Tavel, "What are biomarkers?" *Current Opinion HIV AIDS*, vol. 5, no. 6, pp. 463–466, Nov. 2010.

[5] D. J. Hunter, E. Losina, A. Guermazi, D. Burstein, M. N. Lassere, and V. Kraus, "A pathway and approach to biomarker validation and qualification for osteoarthritis clinical trials," *Current Drug Targets*, vol. 11, no. 5, pp. 536–545, May 2010.

[6] H. Quezada, A. L. Guzmán-Ortiz, H. Díaz-Sánchez, R. Valle-Rios, and J. Aguirre-Hernández, "Omics-based biomarkers: Current status and potential use in the clinic," *Boletín Médico del Hospital Infantil de México*, vol. 74, no. 3, pp. 219–226, May 2017.

[7] X. Guo et al., "CNSA: A data repository for archiving omics data," *Database, J. Biol. Databases Curation*, vol. 2020, p. baaa055, Jan. 2020, doi: 10.1093/database/baaa055.

[8] K. Raja, M. Patrick, Y. Gao, D. Madu, Y. Yang, and L. C. Tsoi, "A review of recent advancement in integrating omics data with literature mining towards biomedical discoveries," *Int. J. Genomics*, vol. 2017, pp. 1–10, 2017.

[9] A. J. Vargas and C. C. Harris, "Biomarker development in the precision medicine era: Lung cancer as a case study," *Nature Rev. Cancer*, vol. 16, no. 8, pp. 525–537, Aug. 2016.

[10] J. Shendure, S. Balasubramanian, G. M. Church, W. Gilbert, J. Rogers, J. A. Schloss, and R. H. Waterston, "DNA sequencing at 40: Past, present and future," *Nature*, vol. 550, no. 7676, pp. 345–353, Oct. 2017.

[11] E. R. Mardis, "DNA sequencing technologies: 2006–2016," *Nature Protocols*, vol. 12, no. 2, pp. 213–218, Feb. 2017.

[12] B. E. Slatko, A. F. Gardner, and F. M. Ausubel, "Overview of next-generation sequencing technologies," *Current Protocols Mol. Biol.*, vol. 122, no. 1, 2018, Art. no. e59.

[13] J. Davis-Turak, S. M. Courtney, E. S. Hazard, W. B. Glen, W. A. da Silveira, T. Wesselman, L. P. Harbin, B. J. Wolf, D. Chung, and G. Hardiman, "Genomics pipelines and data integration: Challenges and opportunities in the research setting," *Expert Rev. Mol. Diag.*, vol. 17, no. 3, pp. 225–237, Mar. 2017.

[14] O. U. Sezerman, E. Ulgen, N. Seymen, and I. M. Durasi, "Bioinformatics Workflows for genomic variant discovery, interpretation and prioritization," in *Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations*, A. Samadikuchaksaraei and M. Seifi, Eds. Rijeka, Croatia: IntechOpen, 2019, ch. 2, doi: 10.5772/intechopen.85524.

[15] N. Ahmed, K. Bertels, and Z. Al-Ars, "A comparison of seed-and-extend techniques in modern DNA read alignment algorithms," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2016, pp. 1421–1428.

[16] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," 2013, *arXiv:1303.3997*.

[17] B. Langmead, C. Wilks, V. Antonescu, and R. Charles, "Scaling read aligners to hundreds of threads on general-purpose processors," *Bioinformatics*, vol. 35, no. 3, pp. 421–432, Feb. 2019.

[18] G. G. Faust and I. M. Hall, "SAMBLASTER: Fast duplicate marking and structural variant read extraction," *Bioinformatics*, vol. 30, no. 17, pp. 2503–2505, Sep. 2014.

[19] Y. W. Yu, D. Yorukoglu, J. Peng, and B. Berger, "Quality score compression improves genotyping accuracy," *Nature Biotechnol.*, vol. 33, no. 3, pp. 240–243, Mar. 2015.

[20] R. Simon and S. Roychowdhury, "Implementing personalized cancer genomics in clinical trials," *Nature Rev. Drug Discovery*, vol. 12, no. 5, pp. 358–369, Apr. 2013.

[21] A. Deshpande, W. Lang, T. McDowell, S. Sivakumar, J. Zhang, J. Wang, F. A. San Lucas, J. Fowler, H. Kadara, and P. Scheet, "Strategies for identification of somatic variants using the ion torrent deep targeted sequencing platform," *BMC Bioinf.*, vol. 19, no. 1, Dec. 2018.

[22] D. C. Koboldt, "Best practices for variant calling in clinical sequencing," *Genome Med.*, vol. 12, no. 1, Dec. 2020, Art. no. 91.

[23] D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, and L. Lichtenstein, "Calling somatic SNVs and indels with Mutect2," *bioRxiv*, p. 861054, Apr. 2023. [Online]. Available: https://www.biorxiv.org/content/10.1101/861054v1, doi: 10.1101/861054.

[24] R. Poplin, P.-C. Chang, D. Alexander, S. Schwartz, T. Colthurst, A. Ku, D. Newburger, J. Dijamco, N. Nguyen, P. T. Afshar, S. S. Gross, L. Dorfman, C. Y. McLean, and M. A. DePristo, "A universal SNP and small-indel variant caller using deep neural networks," *Nature Biotechnol.*, vol. 36, no. 10, pp. 983–987, Nov. 2018.

[25] D. Raimondi, A. Passemiers, P. Fariselli, and Y. Moreau, "Current cancer driver variant predictors learn to recognize driver genes instead of functional variants," *BMC Biol.*, vol. 19, no. 1, pp. 1–12, Jan. 2021.

[26] A.-L. Brown, M. Li, A. Goncearenco, and A. R. Panchenko, "Finding driver mutations in cancer: Elucidating the role of background mutational processes," *PLOS Comput. Biol.*, vol. 15, no. 4, Apr. 2019, Art. no. e1006981.

[27] M. M. Li, M. Datto, E. J. Duncavage, S. Kulkarni, N. I. Lindeman, S. Roy, A. M. Tsimberidou, and C. L. Vnencak-Jones, "Standards and guidelines for the interpretation and reporting of sequence variants in cancer: A joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists," *J. Mol. Diagnostics*, vol. 19, no. 1, pp. 4–23, 2017.

[28] J. J. Shen, J. J. Shen, S. B. Wortmann, L. de Boer, L. A. J. Kluijtmans, M. C. D. G. Huigen, J. Koch, S. Ross, C. D. Collins, R. van der Lee, C. D. M. van Karnebeek, and M. R. Hegde, "The role of clinical response to treatment in determining pathogenicity of genomic variants," *Genet. Med.*, vol. 23, pp. 581–585, Oct. 2020.

[29] Y. Xie, G. Gao, and X. A. Chen, "Outlining the design space of explainable intelligent systems for medical diagnosis," 2019, *arXiv:1902.06019*.

[30] E. Tjoa and C. Guan, ‘‘A survey on explainable artificial intelligence (XAI): Toward medical XAI,’’ *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[31] K. Wang, M. Li, and H. Hakonarson, ‘‘ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data,’’ *Nucleic Acids Res.*, vol. 38, no. 16, Sep. 2010, Art. no. e164.

[32] X. Liu, X. Jian, and E. Boerwinkle, ‘‘dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions,’’ *Hum. Mutation*, vol. 32, pp. 894–899, Aug. 2011.

[33] M. Mesbah-Uddin, ‘‘Prediction of deleterious nonsynonymous SNPs by integrating multiple classifiers—An application to neurodegenerative diseases,’’ *PeerJ*, Apr. 2023, Art. no. e1224. [Online]. Available: https://peerj.com/preprints/994, doi: 10.7287/peerj.preprints.994v1.

[34] D. Anderson and T. Lassmann, ‘‘A phenotype centric benchmark of variant prioritisation tools,’’ *NPJ Genome Med.*, vol. 3, no. 1, p. 5, 2018.

[35] V. Suybeng, F. Koeppel, A. Harlé, and E. Rouleau, ‘‘Comparison of pathogenicity prediction tools on somatic variants,’’ *J. Mol. Diag.*, vol. 22, no. 12, pp. 1383–1392, 2020.

[36] P. K. S. Ng et al., ‘‘Systematic functional annotation of somatic mutations in cancer,’’ *Cancer Cell*, vol. 33, pp. 450–462, Mar. 2018.

[37] M. C. Benton, R. A. Smith, L. M. Haupt, H. G. Sutherland, P. J. Dunn, C. L. Albury, N. Maksemous, R. Lea, and L. Griffiths, ‘‘Variant call format–diagnostic annotation and reporting tool: A customizable analysis pipeline for identification of clinically relevant genetic variants in next-generation sequencing data,’’ *J. Mol. Diag.*, vol. 21, no. 6, pp. 951–960, 2019.

[38] D. Ruden, V. M. Patel, M. Coon, T. Nguyen, S. J. Land, D. M. Ruden, and X. Lu, ‘‘Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift,’’ *Frontiers Genet.*, vol. 3, p. 35, Mar. 2012.

[39] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, ‘‘The ensembl variant effect predictor,’’ *Genome Biol.*, vol. 17, no. 1, p. 122, Apr. 2023, doi: 10.1186/s13059-016-0974-4.

[40] Y. Mao, H. Chen, H. Liang, F. Meric-Bernstam, G. B. Mills, and K. Chen, ‘‘CanDrA: Cancer-specific driver missense mutation annotation with optimized features,’’ *PLoS ONE*, vol. 8, no. 10, 2013, Art. no. e77945.

[41] K. Tomczak, P. Czerwinska, and M. Wiznerowicz, ‘‘The cancer genome atlas (TCGA): An immeasurable source of knowledge,’’ *Contemp. Oncol.*, vol. 19, no. 1A, pp. A68–A77, 2015.

[42] N. Bindal, S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, C. Y. Kok, M. Jia, S. Bamford, C. Cole, S. Ward, J. Teague, M. R. Stratton, P. Campbell, and A. P Futreal, ‘‘COSMIC: The catalogue of somatic mutations in cancer,’’ *Nucleic Acids Res.*, vol. 47, no. D1, pp. D941–D947, 2018.

[43] M. Kircher, D. M. Witten, P. Jain, B. J. O’Roak, G. M. Cooper, and J. A. Shendure, ‘‘A general framework for estimating the relative pathogenicity of human genetic variants,’’ *Nature Genet.*, vol. 46, pp. 310–315, Feb. 2014.

[44] D. Quang, Y. Chen, and X. Xie, ‘‘DANN: A deep learning approach for annotating the pathogenicity of genetic variants,’’ *Bioinformatics*, vol. 31, no. 5, pp. 761–763, 2015.

[45] B. J. Feng, ‘‘PERCH: A unified framework for disease gene prioritization,’’ *Hum. Mutation*, vol. 38, no. 3, pp. 243–251, 2017.

[46] D. Raimondi, I. Tanyalcin, J. Ferté, A. Gazzo, G. Orlando, T. Lenaerts, M. Rooman, and W. Vranken, ‘‘DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins,’’ *Nucleic Acids Res.*, vol. 45, pp. W201–W206, Jul. 2017.

[47] N. Alirezaie, K. D. Kernohan, T. Hartley, J. Majewski, and T. Hocking, ‘‘ClinPred: Prediction tool to identify disease-relevant nonsynonymous single-nucleotide variants,’’ *Amer. J. Hum. Genet.*, vol. 103, no. 4, pp. 474–483, 2018.

[48] C. Tokheim and R. Karchin, ‘‘Chasmplus reveals the scope of somatic missense mutations driving human cancers,’’ *Cell Syst.*, vol. 9, no. 1, pp. 9–23, 2019.

[49] M. F. Rogers, H. A. Shihab, M. E. Mort, D. N. Cooper, T. R. Gaunt, and C. Campbell, ‘‘FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features,’’ *Bioinformatics*, vol. 34, no. 3, pp. 511–513, 2018.

[50] C. Li, D. Zhi, K. Wang, and X. Liu, ‘‘MetaRNN: differentiating rare pathogenic and rare benign missense SNVs and InDels using deep learning,’’ *Genome Med.*, vol. 14, no. 1, p. 115, Apr. 2023, doi: 10.1186/s13073-022-01120-z.

[51] H. Wang, T. Wang, X. Zhao, H. Wu, M. You, Z. Sun, and F. Mao, ‘‘AI-Driver: An ensemble method for identifying driver mutations in personal cancer genomes,’’ *NAR Genomics Bioinf.*, vol. 2, no. 4, 2020, Art. no. lqaa084.

[52] D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, A. Rovira, I. Tusquets, J. Albanell, J. Rodon, J. Tabernero, C. de Torres, R. Dienstmann, A. Gonzalez-Perez, and N. Lopez-Bigas, ‘‘Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations,’’ *Genome Med.*, vol. 10, no. 1, 2018, Art. no. 25.

[53] J. Zhang, R. Bajari, D. Andric, F. Gerthoffert, A. Lepsa, H. Nahal-Bose, L. D. Stein, and V. Ferretti, ‘‘The international cancer genome consortium data portal,’’ *Nature Biotechnol.*, vol. 37, pp. 367–369, Mar. 2019.

[54] A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, D. Tamborero, M. P. Schroeder, A. Jene-Sanz, A. Santos, and N. Lopez-Bigas, ‘‘IntOGen-mutations identifies cancer drivers across tumor types,’’ *Nature Methods*, vol. 10, no. 11, pp. 1081–1082, Nov. 2013.

[55] F. Martínez-Jiménez, F. Muiños, I. Sentís, J. Deu-Pons, I. Reyes-Salazar, C. Arnedo-Pac, L. Mularoni, O. Pich, J. Bonet, H. Kranas, A. Gonzalez-Perez, and N. Lopez-Bigas, ‘‘A compendium of mutational cancer driver genes,’’ *Nature Rev. Cancer*, vol. 20, pp. 555–572, Aug. 2020.

[56] X. Shi, H. Teng, L. Shi, W. Bi, W. Wei, F. Mao, and Z. Sun, ‘‘Comprehensive evaluation of computational methods for predicting cancer driver genes,’’ *Briefings Bioinf.*, vol. 23, no. 2, 2022, Art. no. bbab548.

[57] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and K. Sirotkin, ‘‘dbSNP: The NCBI database of genetic variation,’’ *Nucleic acids Res.*, vol. 29, pp. 308–311, Jan. 2001.

[58] M. Olivier, ‘‘A haplotype map of the human genome,’’ *Nature*, vol. 437, no. 7063, pp. 1299–1320, Oct. 2005.

[59] H. Li, ‘‘A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data,’’ *Bioinformatics*, vol. 27, no. 21, pp. 2987–2993, Nov. 2011.

[60] K. J. Karczewski et al., ‘‘The mutational constraint spectrum quantified from variation in 141,456 humans,’’ *Nature*, vol. 581, pp. 434–443, May 2020.

[61] *IntOGen Documentation Release 1.0*, BBGLab, Barcelona, Spain, 2019.

[62] D. L. Wheeler, ‘‘Database resources of the national center for biotechnology information,’’ *Nucleic Acids Res.*, vol. 33, pp. D39–D45, Dec. 2004.

[63] W. J. Kent, ‘‘The human genome browser at UCSC,’’ *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002.

[64] M. Lek et al., ‘‘Analysis of protein-coding genetic variation in 60,706 humans,’’ *Nature*, vol. 536, pp. 285–291, Aug. 2016.

[65] X. Liu, C. Li, C. Mou, Y. Dong, and Y. Tu, ‘‘dbNSFP v4: A comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs,’’ *Genome Med.*, vol. 12, no. 1, Dec. 2020, Art. no. 103.

[66] M. J. Landrum, J. M. Lee, G. R. Riley, W. Jang, W. S. Rubinstein, D. M. Church, and D. R. Maglott, ‘‘ClinVar: Public archive of relationships among sequence variation and human phenotype,’’ *Nucleic Acids Res.*, vol. 42, no. D1, pp. D980–D985, Jan. 2014.

[67] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, and H. L. Rehm, ‘‘Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology,’’ *Genet. Med.*, vol. 17, no. 5, pp. 405–424, May 2015.

[68] G. Glusman, J. Caballero, D. E. Mauldin, L. Hood, and J. C. Roach, ‘‘Kaviar: An accessible system for testing SNV novelty,’’ *Bioinformatics*, vol. 27, no. 22, pp. 3216–3217, Nov. 2011.

[69] S. McCarthy et al., ‘‘A reference panel of 64,976 haplotypes for genotype imputation,’’ *Nature Genet.*, vol. 48, no. 10, pp. 1279–1283, Oct. 2016, doi: 10.1038/ng.3643.

[70] 1000 Genomes Project Consortium, ‘‘A global reference for human genetic variation,’’ *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[71] P. L. Auer et al., ‘‘Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO exome sequencing project,’’ *Amer. J. Hum. Genet.*, vol. 91, no. 5, pp. 794–808, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0002929712004739, doi: 10.1016/j.ajhg.2012.08.031.

[72] H. Qi, H. Zhang, Y. Zhao, C. Chen, J. J. Long, W. K. Chung, Y. Guan, and Y. Shen, ‘‘MVP predicts the pathogenicity of missense variants by deep learning,’’ *Nature Commun.*, vol. 12, no. 1, Jan. 2021, Art. no. 510.

[73] L. Sundaram, H. Gao, S. R. Padigepati, J. F. McRae, Y. Li, J. A. Kosmicki, N. Fritzilas, J. Hakenberg, A. Dutta, J. Shon, J. Xu, S. Batzoglou, X. Li, and K. K.-H. Farh, ‘‘Predicting the clinical impact of human mutation with deep neural networks,’’ *Nature Genet.*, vol. 50, no. 8, pp. 1161–1170, Aug. 2018.

[74] X. Liu, C. Wu, C. Li, and E. Boerwinkle, "DbNSFP v3.0: A one-stop database of functional predictions and annotations for human non-synonymous and splice-site SNVs," *Human Mutation*, vol. 37, no. 3, pp. 235–241, Mar. 2016.

[75] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLoS Comput. Biol.*, vol. 6, no. 12, Dec. 2010, Art. no. e1001025.

[76] S. Chun and J. C. Fay, "Identification of deleterious mutations within three human genomes," *Genome Res.*, vol. 19, no. 9, pp. 1553–1561, Sep. 2009.

[77] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res.*, vol. 20, no. 1, pp. 110–121, Jan. 2010.

[78] J. Felsenstein and G. A. Churchill, "A hidden Markov model approach to variation among sites in rate of evolution," *Mol. Biol. Evol.*, vol. 13, no. 1, pp. 93–104, Jan. 1996.

[79] M. Garber, M. Guttman, M. Clamp, M. C. Zody, N. Friedman, and X. Xie, "Identifying novel constrained elements by exploiting biased substitution patterns," *Bioinformatics*, vol. 25, no. 12, pp. i54–i62, Jun. 2009.

[80] P. C. Ng and S. Henikoff, "SIFT: Predicting amino acid changes that affect protein function," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3812–3814, Jul. 2003.

[81] R. Steinhaus, S. Proft, M. Schuelke, D. N. Cooper, J. M. Schwarz, and D. Seelow, "MutationTaster2021," *Nucleic Acids Res.*, vol. 49, no. W1, pp. W446–W451, Jul. 2021.

[82] B. Reva, Y. Antipin, and C. Sander, "Predicting the functional impact of protein mutations: Application to cancer genomics," *Nucleic Acids Res.*, vol. 39, no. 17, Sep. 2011, Art. no. e118.

[83] B. Reva, Y. Antipin, and C. Sander, "Determinants of protein function revealed by combinatorial entropy optimization," *Genome Biol.*, vol. 8, no. 11, p. R232, 2007.

[84] H. A. Shihab, J. Gough, M. Mort, D. N. Cooper, I. N. Day, and T. R. Gaunt, "Ranking non-synonymous single nucleotide polymorphisms based on disease concepts," *Hum. Genomics*, vol. 8, no. 1, pp. 11–16, Dec. 2014.

[85] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the functional effect of amino acid substitutions and indels," *PLoS ONE*, vol. 7, no. 10, Oct. 2012, Art. no. e46688.

[86] Y. Choi, "A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein," in *Proc. ACM Conf. Bioinf., Comput. Biol. Biomed.*, Oct. 2012, pp. 414–417.

[87] H. Carter, C. Douville, P. D. Stenson, D. N. Cooper, and R. Karchin, "Identifying Mendelian disease genes with the variant effect scoring tool," *BMC Genomics*, vol. 14, no. S3, May 2013.

[88] N. Malhis, M. Jacobson, S. J. M. Jones, and J. Gsponer, "LIST-s2: Taxonomy based sorting of deleterious missense mutations across species," *Nucleic Acids Res.*, vol. 48, no. W1, pp. W154–W161, Jul. 2020.

[89] K. E. Samocha, J. A. Kosmicki, K. J. Karczewski, A. H. O'Donnell-Luria, E. Pierce-Hoffman, D. G. MacArthur, B. M. Neale, and M. J. Daly, "Regional missense constraint improves variant deleteriousness prediction," *bioRxiv*, p. 148353, Apr. 2023. [Online]. Available: https://www.biorxiv.org/content/10.1101/148353v1, doi: 10.1101/148353.

[90] B. Gulko, M. J. Hubisz, I. Gronau, and A. C. Siepel, "Probabilities of fitness consequences for point mutations across the human genome," *Nature Genet.*, vol. 47, no. 3, pp. 276–283, 2015.

[91] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature Methods*, vol. 7, no. 4, pp. 248–249, 2010.

[92] B. Li, V. G. Krishnan, M. E. Mort, F. Xin, K. K. Kamati, D. N. Cooper, S. D. Mooney, and P. Radivojac, "Automated inference of molecular mechanisms of disease from amino acid substitutions," *Bioinformatics*, vol. 25, no. 21, pp. 2744–2750, Nov. 2009.

[93] S. Kim, J.-H. Jhong, J. Lee, and J.-Y. Koo, "Meta-analytic support vector machine for integrating multiple omics data," *BioData Mining*, vol. 10, no. 1, Dec. 2017, Art. no. 2.

[94] N. M. Ioannidis et al., "REVEL: An ensemble method for predicting the pathogenicity of rare missense variants," *Amer. J. Hum. Genet.*, vol. 99, no. 4, pp. 877–885, 2016.

[95] K. A. Jagadeesh, A. M. Wenger, M. J. Berger, H. Guturu, P. D. Stenson, D. N. Cooper, J. A. Bernstein, and G. Bejerano, "M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity," *Nature Genet.*, vol. 48, no. 12, pp. 1581–1586, Dec. 2016.

[96] H. A. Shihab, M. F. Rogers, J. Gough, M. Mort, D. N. Cooper, I. N. M. Day, T. R. Gaunt, and C. Campbell, "An integrative approach to predicting the functional effects of non-coding and coding sequence variation," *Bioinformatics*, vol. 31, no. 10, pp. 1536–1543, May 2015.

[97] I. Ionita-Laza, K. McCallum, B. Xu, and J. D. Buxbaum, "A spectral approach integrating functional genomic annotations for coding and noncoding variants," *Nature Genet.*, vol. 48, no. 2, pp. 214–220, Feb. 2016.

[98] Q. Lu, Y. Hu, J. Sun, Y. Cheng, K.-H. Cheung, and H. Zhao, "A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data," *Sci. Rep.*, vol. 5, no. 1, pp. 1–13, May 2015.

[99] Y.-F. Huang, B. Gulko, and A. Siepel, "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data," *Nature Genet.*, vol. 49, no. 4, pp. 618–624, Apr. 2017, doi: 10.1038/ng.3810.

[100] S. L. Harris and D. M. Harris, *Digital Design and Computer Architecture*, 2nd ed. San Mateo, CA, USA: Morgan Kaufmann, 2012.

[101] S. Balasubramanian, Y. Fu, M. Pawashe, P. McGillivray, M. Jin, J. Liu, K. J. Karczewski, D. G. MacArthur, and M. Gerstein, "Using ALoFT to determine the impact of putative loss-of-function variants in protein-coding genes," *Nature Commun.*, vol. 8, Aug. 2017, Art. no. 382.

[102] L. J. Carithers et al., "A novel approach to high-quality postmortem tissue procurement: The GTEx project," *Biopreservation Biobanking*, vol. 13, no. 5, pp. 311–319, 2015.

[103] M. Blum et al., "The InterPro protein families and domains database: 20 years on," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D344–D354, 2020.

[104] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging Technology in Modelling and Graphics* (Advances in Intelligent Systems and Computing), J. K. Mandal and D. Bhattacharya, Eds. Singapore: Springer, 2020, pp. 99–111, doi: 10.1007/978-981-13-7403-6_11.

[105] M. Lin, Q. Chen, and S. Yan, "Network in network," *CoRR*, vol. abs/1312.4400, pp. 1–10, Dec. 2014.

[106] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[107] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI*, 2017, pp. 1–7.

[108] G. Luo, "A review of automatic selection methods for machine learning algorithms and hyper-parameter values," *Netw. Model. Anal. Health Informat. Bioinf.*, vol. 5, pp. 1–16, May 2016.

[109] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," 2020, *arXiv:2003.05689*.

[110] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," 2020, *arXiv:2007.15745*.

[111] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[112] L. Li, K. Jamieson, A. Rostamizadeh, E. Gonina, M. Hardt, B. Recht, and A. Talwalkar, "A system for massively parallel hyperparameter tuning," 2018, *arXiv:1810.05934*.

[113] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[114] Ž. Vujović, "Classification model evaluation metrics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 599–606, 2021.

[115] R. Dias and A. Torkamani, "Artificial intelligence in clinical and genomic diagnostics," *Genome Med.*, vol. 11, no. 1, pp. 1–12, Dec. 2019.

[116] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[117] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *J. Comput. Graph. Statist.*, vol. 24, no. 1, pp. 44–65, 2013.

[118] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.

[119] A. Nazábal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using VAEs," *Pattern Recognit.*, vol. 107, Nov. 2020, Art. no. 107501.

[120] P.-A. Mattei and J. Frellsen, "MIWAE: Deep generative modelling and imputation of incomplete data sets," in *Proc. ICML*, 2019, pp. 4413–4423.

MATTEO BASTICO (Student Member, IEEE) received the B.S. degree in information engineering from Universitã degli Studi di Padova (UniPD), Italy, in 2018, the M.S. degree in telecommunication engineering from Universidad Politécnica de Madrid (UPM), Spain, in 2021, and the M.S. degree (cum laude) in ICT for internet and multimedia from UniPD, in 2021. Since 2021, he has been a Researcher in artificial intelligence and deep learning with the Visual Telecommunications Applications Group (GATV), UPM. He has been focused on health and sensors scenarios developing technical solutions in EU projects.

ANAIDA FERNÁNDEZ-GARCÍA (Student Member, IEEE) received the bachelor's degree in sound and image in telecommunication engineering from Universidad de Alicante (UA), Spain, in 2020. She is currently pursuing the Master of Science degree in telecommunication engineering with Universidad Poltécnica de Madrid (UPM), Spain. She is also collaborating with the Visual Telecommunications Applications Group (GATV), UPM, for her Master thesis on artificial intelligence applied on genomics and radiomics.

ALBERTO BELMONTE-HERNÁNDEZ (Member, IEEE) received the degree in telecommunication engineering, the master's degree with a focus on communication systems, and the Ph.D. degree (cum laude) from Universidad Politécnica de Madrid (UPM), in 2014, 2016, and 2020, respectively. Currently, he is an assistant professor in several subjects. Since 2016, he has been with the Visual Telecommunications Applications Group (GATV), UPM. He is actively working on artificial intelligent applied to multimedia content and sensors for pattern detection, recognition, and fusion. He has been developing technical parts in national and EU projects.

SILVIA URIBE MAYORAL received the degree (Hons.) in telecommunication engineering, the master's degree in communications technologies and systems, the master's degree in telecommunication management, and the Ph.D. degree (cum laude) from Universidad Politécnica de Madrid, in February 2008, September 2010, 2013, and 2016, respectively. She has been a member of the Visual Telecommunications Applications Group (GATV), UPM, since 2006. Her research interests include interactivity technologies, content personalization technologies, and big data. Related to this, she has been participating with technical responsibilities in some national and European projects, and she is the author and coauthor of several papers and scientific contributions in international conferences and journals.