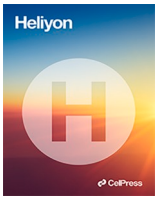
Heliyon 10 (2024) e39379

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Predicting somatic mutation origins in cell-free DNA by semi-supervised GAN models

Fahimeh Palizban [a], Mohammadmahdi Sarbishegi [b], Kaveh Kavousi [a,**], Mahya Mehrmohamadi [b,*]

[a] *Laboratory of Complex Biological Systems and Bioinformatics (CBB), Institute of Biochemistry and Biophysics (IBB), University of Tehran, Tehran, Iran*

[b] *Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran*

ARTICLE INFO

Keywords:
Somatic mutation
Cancer
Cell free DNA (cfDNA)
Genomics
Clonal hematopoiesis (CH)
Semi-supervised learning
Semi-supervised generative adversarial network

ABSTRACT

Motivation: Distinguishing between pathogenic cancer-associated mutations and other somatic variants present in cell-free DNA (cfDNA) is one of the challenges in the field of liquid biopsy. This distinction is critical, since the misclassification of mutations stemming from clonal hematopoiesis (CH) as tumor-derived and vice versa could result in inaccurate diagnoses and inappropriate therapeutic interventions for patients.

Results: We addressed this by developing a specialized machine learning technique to differentiate tumor- or CH-related mutations in cfDNA. We established a comprehensive in-house reference catalog, comprising approximately 25,000 single nucleotide variants (SNVs), each linked to either tumor or CH origin. This reference serves as a foundation for training a deep learning model, which is structured on the semi-supervised generative adversarial network (SSGAN) architecture. By analyzing genomic coordinates and nucleotide composition of cfDNA variants, our model attains 95 % area under the curve (AUC) in classifying uncharacterized variants as CH or tumor-derived. In conclusion, our research emphasizes the potential of genomic feature prediction, using cfDNA data, to stand as a robust alternative to conventional multi-analyte sequencing methods. This approach not only enhances the accuracy of distinguishing CH from tumor mutations in liquid biopsy data, but also highlights the potential of advanced data analysis techniques and machine learning in genomics and personalized medicine. **Availability**: https://github.com/FPalizban/SSGAN.

1. Introduction

Liquid biopsy testing relies on the identification of cancer biomarkers within bodily fluids. These biomarkers encompass genetic and epigenetic changes linked to cancer [1]. cfDNA is particularly important for clinical applications due to its advantages, such as its minimally invasive collection and its ability to capture molecular alterations in tumors [1]. Nevertheless, somatic mosaicism in plasma complicates the accurate interpretation of liquid biopsy results [2]. A substantial portion of current tests for cancer monitoring and

* Corresponding author.
** Corresponding author.
E-mail addresses: fahimehpalizban@ut.ac.ir (F. Palizban), mahdi.sarbishegi@ut.ac.ir (M. Sarbishegi), kkavousi@ut.ac.ir (K. Kavousi), mehrmohamadi@ut.ac.ir (M. Mehrmohamadi).

https://doi.org/10.1016/j.heliyon.2024.e39379
Received 27 June 2024; Received in revised form 12 October 2024; Accepted 13 October 2024
Available online 15 October 2024

early detection leverages somatic mutations as markers of malignancy. However, in recent years, the cancer-specificity of somatic mutations in cfDNA has come under scrutiny, given the emerging realization that non-cancer-related events within the body can also give rise to somatic mutations [3]. Clonal hematopoiesis (CH) is a biological phenomenon resulting from the expansion of white blood cells originating from a single hematopoietic stem cell. It is often driven by mutations in genes that confer a competitive advantage to hematopoietic stem cells (HSCs), particularly under certain physiological conditions [4]. CH is a natural facet of the aging process, characterized by the accumulation of somatic mutations and the clonal expansion of hematopoietic stem cells [5]. Non-tumor-derived CH mutations pose a challenge in analyzing liquid biopsies, introducing 'background noise' that hinders accurate diagnosis. Misclassifying CH mutations as tumor-derived somatic mutations may result in erroneous diagnoses and inappropriate therapeutic interventions [6]. A study deploying a highly sensitive and specific circulating tumor DNA (ctDNA) sequencing assay uncovered that approximately 53.2 % of mutations in cancer patients exhibited features consistent with clonal hematopoiesis [7]. Researchers have studied different types of somatic mutations in blood to determine their origins. This typically requires matched sequencing of circulating cell-free DNA (cfDNA) and white blood cells (WBCs). In a recent study, a statistical model demonstrated high accuracy in categorizing these two types of variants [8]. Another study involved the analysis of blood whole-exome sequencing (WES) data from a cohort of 200,453 participants within the UK Biobank. This comprehensive analysis unveiled 43 genes harboring somatic mutations associated with clonal hematopoiesis, providing valuable insights into the genetic factors underpinning this phenomenon [9]. Another study used the IntOGen pipeline to identify new genes associated with clonal hematopoiesis [10]. Recent studies in this domain have largely been focused on refining experimental methodologies to eliminate background noise within cfDNA sequencing data, primarily arising from clonal hematopoiesis and mitochondrial mutations. For instance, parallel sequencing of cfDNA and WBCs was carried out in a cohort exceeding 10,000 Chinese patients. This involved the implementation of various statistical filtration steps, culminating in the identification of candidate driver genes in different cancer types [11]. To closely monitor therapeutic dynamics, another group of researchers conducted an in-depth investigation into the mechanisms of the transition from normal hematopoietic stem and progenitor cells (HSPCs) to cells characterized by significantly enhanced proliferation rates [12]. Prior investigations have traditionally relied on the integration of WBC sequencing data, cfDNA, and matched tumor data to discern various somatic mutation types. As in a recent

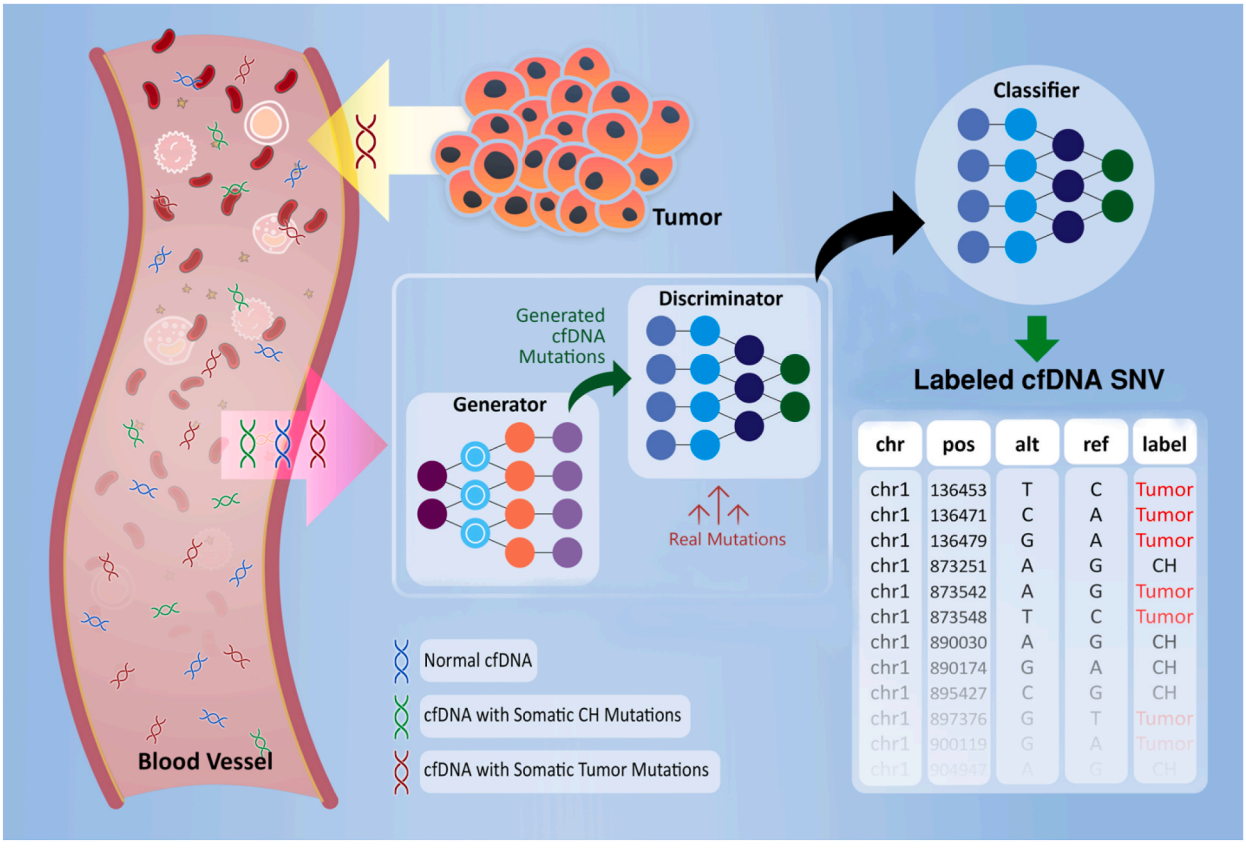


Fig. 1. The Schematic overview of the cfDNA SNV type prediction by the SSGAN model involves a multi-component process. The generator takes noise as input and employs a deep neural network. It then generates synthetic cfDNA SNVs based on the learned distributions. The discriminator, on the other hand, takes both real unlabeled cfDNA SNVs and simulated cfDNA SNVs as input and employs a classification model again based on deep neural network. The training data of GAN, comprising real cfDNA SNVs without known labels. The output of the classifier is a labeled cfDNA SNVs, representing the predicted labels for both CH or tumor SNVs based on the locally developed somatic SNV catalog. The prebuilt model was trained based on prostate cancer dataset and validated with glioma dataset as well.

study, a tool named PLASMUT was developed to measure the probability of mutation in cfDNA as tumor specific or not. This approach requires the both plasma and WBCs samples along with the corresponding number of total distinct reads at those locations [13]. However, the evolving landscape of machine learning techniques presents an opportunity to propose more efficient and cost-effective alternatives for analyzing biological data. To this end, a previous study employed a supervised learning approach, leveraging multiple genomic and functional attributes within a labeled variant dataset to differentiate between clonal CH and tumor-derived mutations in cfDNA samples. The study, however, encountered constraints due to the limited number of variants at its disposal, attributed to the scarcity of validated labeled datasets in this domain [14].

In summary, current methods, which often require extensive resources, are inadequate due to their high cost and time consumption. In the present study, we harnessed machine learning methodologies within the family of generative adversarial networks (GAN) algorithms, specifically for semi-supervised learning [15]. Our primary aim was to categorize raw variants derived from cfDNA genomic data into two fundamental groups of CH or tumor-derived somatic variants. Our study introduces a novel semi-supervised GAN (SSGAN) model designed to improve the precision of identifying the origins of cfDNA somatic mutations, addressing these limitations effectively. Generative Adversarial Networks (GANs) are particularly suitable for this task due to their unique architecture, which consists of a generator and a discriminator. The generator creates synthetic data that mimics the real data distribution, while the discriminator evaluates whether the data is real or generated. This adversarial process helps the model learn complex patterns and hidden aspects of the data, which is particularly useful when dealing with the high-dimensional and intricate nature of genomic data. By capturing the underlying distribution of the data, GANs can reveal hidden structures and relationships that may not be immediately apparent through traditional methods. The semi-supervised learning (SSL) approach enhances this capability by utilizing both labeled and unlabeled data during training. In the context of cfDNA, obtaining a comprehensive set of labeled data is often challenging due to the costs and efforts associated with manual annotation. However, there is typically a larger pool of unlabeled data available. By integrating GAN with SSL, the model can leverage the structure and distribution of the unlabeled data to improve learning and generalization. This dual advantage makes the SSGAN model a powerful tool for distinguishing between CH and tumor-derived mutations, offering a cost-effective and highly accurate alternative. An illustrative depiction of our study is shown in Fig. 1. The integration of GANs in our methodology allows for the generation of realistic synthetic data that augments the training process, enabling the model to learn from a more extensive dataset than would otherwise be possible. This capability is crucial in scenarios where labeled data is scarce but unlabeled data is abundant, as is often the case with cfDNA samples.

2. Results

2.1. Preparing a comprehensive set of somatic SNVs

While some mutations that drive CH have been identified through experimental and epidemiological studies, a comprehensive compilation of all genes that may contribute to CH-related mutations in hematopoietic stem cells (HSCs) remains incomplete. The success of the training phase in any machine learning-based classifier is relevant to the availability of suitable labeled data. To address this, our initial efforts were directed at collecting lists of somatic mutations known as either CH or tumor-derived from previously published resources. We explored several databases housing information on tumor somatic variations, including The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), Catalogue of Somatic Mutations in Cancer (COSMIC), and Precision Oncology Knowledge Base (OncoKB). Our goal was to assemble a comprehensive catalog of somatic mutations labeled as either tumor-derived or CH. To accomplish this, we curated a comprehensive somatic variant reference including both CH and tumor derived from several studies [7, The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020, 9]. Our final somatic SNV catalog encompasses approximately 25,000 non-redundant SNVs, annotated according to their origin as tumor or CH, based on the studies mentioned above (Supplementary Material 1). Prior to training, common SNVs between the two categories (labeled as both CH and tumor-derived according to the literature) were removed our final set. This SNV catalog served as the training data for the classification task of our proposed SSGAN model. Subsequently, we further annotated the these SNVs to identify the corresponding genes for each CH and tumor group. This analysis revealed a total of 7882 non-redundant genes for the tumor group and 99 for the CH group, with 16 genes shared between the two categories (Supplementary Material 2). It is worth mentioning that although there are overlaps between the two groups of CH and tumor on the gene resolution, the redundancy was eliminated between them on the variant level and each group has its unique SNVs. As illustrated in the Supplementary Material 1, although CH-related mutations are annotated with limited number of genes, they are also distributed throughout the genome covering different chromosomes. This broader distribution helps mitigate the risk of the model focusing solely on specific chromosomes or gene positions. And during the model training, the mutation data was inserted to the model and not the gene name or any other type of variant annotation data. Additionally, to avoid potential overfitting and ensure that the model does not become biased towards a particular category, we carefully balanced the number of mutations from both CH-related and cancer-related categories when training the model. By maintaining an equal number of mutations from each category, we ensured that the model was exposed to a representative sample from both mutation types. This balancing approach is crucial for preventing the model from "memorizing" specific patterns associated with either CH-related or cancer-related mutations and instead encourages the model to learn more generalizable features that can be applied to mutations in various contexts.

2.2. Building a classifier to predict somatic mutation types

After assembling a comprehensive catalog of labeled somatic single nucleotide variants (SNVs), we initiated our study using a prostate cancer dataset that included 48 triplet samples (cfDNA, tissue, and white blood cells) [16]. The SNVs called from the 20 cfDNA

samples in this dataset served as the unlabeled portion for our model input. Given the mixture of labeled and unlabeled SNVs in our dataset, we aimed to develop a deep-learning-based approach capable of predicting mutation types from cfDNA using semi-supervised methods. Generative Adversarial Networks (GANs) have recently shown significant promise in semi-supervised learning (SSL) frameworks across various domains [17]. The strength of GANs lies in their dual network architecture, consisting of a generator and a discriminator, which allows for the effective learning of complex data distributions. Integrating GANs with semi-supervised learning can lead to superior results by leveraging the rich information inherent in both labeled and unlabeled data, thereby uncovering hidden patterns and relationships within the input data. In our approach, we applied GANs within a semi-supervised training context. Specifically, we utilized the discriminator network as a classifier to output the class of each SNV in the classification phase. The process begins with data integration through the generative component of the model, which synthesizes realistic cfDNA SNVs. This is followed by a semi-supervised classification task where the discriminator network is trained to distinguish between different classes of SNVs. Our proposed method produces two main outputs: simulated cfDNA SNVs and classified SNVs from real cfDNA genomic data. This dual output allows us to assess the model's performance in two distinct phases. The first phase evaluates the generative component of the GAN, focusing on the quality and realism of the simulated SNVs. The second phase assesses the performance of the discriminator as a classifier, measured by its ability to accurately classify SNVs in real cfDNA data. The performance of our classifier was tested using the prostate cancer dataset during the initial phase of model development. As depicted in Fig. 2A, the results demonstrate the classifier's capability to accurately predict mutation types, thereby validating our approach. The effectiveness of the model was assessed using loss functions tailored to both the GAN and the classifier, ensuring a comprehensive evaluation of its performance. Here, integrating GANs with semi-supervised learning, not only improves classification accuracy but also enhances the model's ability to generalize from a diverse set of data, addressing the inherent challenges in cfDNA analysis.

Our input features were extracted from the VCF files, encompassing essential information such as chromosome number, SNV position, reference nucleotide, alternate nucleotide, and the nucleotide sequence context around each variant (three nucleotides upstream and downstream). These features were selected to capture both positional and sequence-specific information pertinent to distinguishing between tumor-derived and clonal hematopoiesis (CH)-related somatic variants. No additional annotations were included to keep the feature set focused on primary genomic data. To assess the importance of these features, we employed a method for feature ranking tailored to semi-supervised learning contexts, utilizing ensemble ranking techniques. Ensemble ranking involves aggregating the results of multiple models to provide a robust estimate of feature importance. This method helps to mitigate biases that might arise from a single model and ensures a more reliable ranking of features. It's important to note that GANs are primarily designed for unsupervised representation learning, rather than for feature ranking. The strength of GAN models lies in their ability to learn complex, high-dimensional data distributions without requiring labeled data. Given this, GANs are not typically used to assess the

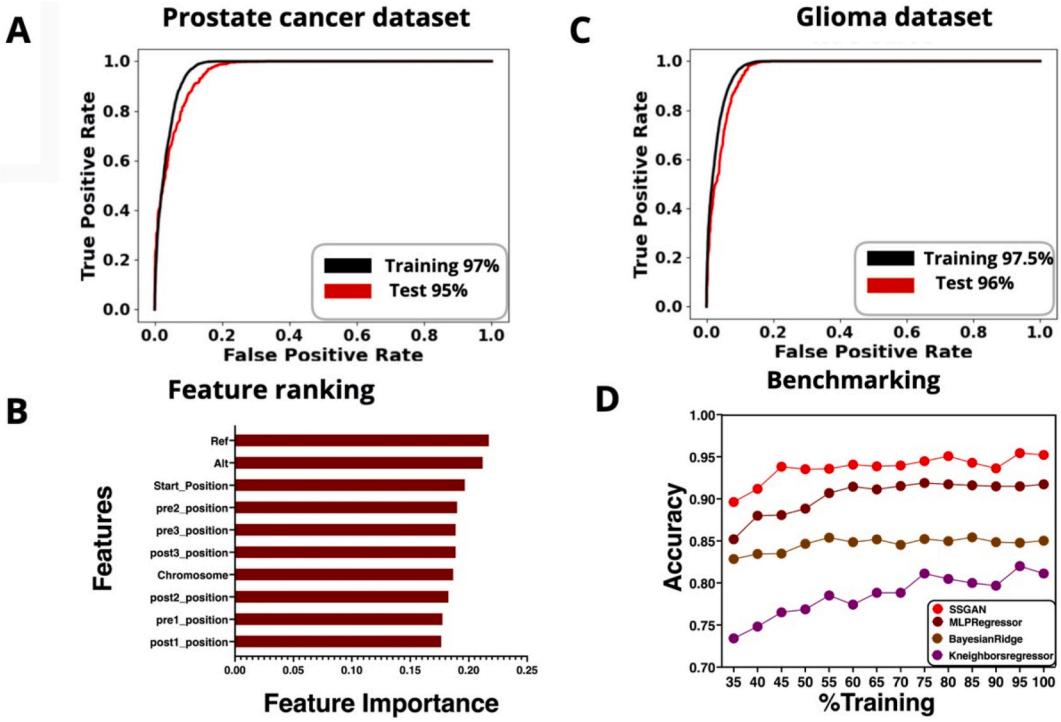


Fig. 2. SSGAN model feature analysis and performance assessment. **A)** The ROC curve showing the performance of SSGAN model based on prostate cancer dataset. **B)** Feature selection and ranking by applying ensemble learning and measuring the importance of each feature. **C)** The ROC curve showing the performance of SSGAN model based on glioma as an independent dataset for further assessment. **D)** Comparing the performance of SSGAN model by running along the different methods on the same prostate cancer dataset based on accuracy metric.

relative importance of individual features in a dataset. And in our study, we used semi-supervised methods based on Ensemble learning for feature ranking instead of relying on the GAN itself for this purpose. The rationale behind this choice is that our ultimate application of SSGAN is differentiating between CH-related and tumor-derived mutations in cfDNAis a semi-supervised task. Therefore, we employed semi-supervised methods to prioritize features such as chromosome number, SNV position, reference and alternate nucleotides, and SNV nucleotide context. These semi-supervised methods provided a robust framework for ranking features, ensuring that the most informative attributes were selected and appropriately weighted during the model training phase. While the surrogate models provided an initial understanding of feature importance, these rankings were not directly imposed on the GAN. Instead, they served as a guide for feature selection, ensuring that the GAN received the most relevant inputs for learning. The feature importance ranking, as illustrated in Fig. 2B, revealed that the nucleotide type at the reference position and its alteration (alternate nucleotide) exhibited the highest importance among all features. This finding is consistent with biological expectations, as the nature of the mutation itself is crucial for distinguishing between different types of somatic variants. The relatively balanced importance of other features, such as chromosome number, SNV position, and surrounding nucleotide sequences, indicates that the model effectively utilizes all provided information, enhancing classification robustness. This holistic approach ensures a comprehensive understanding of the mutation landscape, leveraging the collective importance of diverse features to improve accuracy. The ensemble ranking method used confirms that each feature's contribution is appropriately weighted, aligning with biological expectations and validating our feature selection strategy.

For the assessment of our pre-built deep-learning model, we used a completely distinct dataset from a different cancer type (Glioma SRP2668702). The pre-built model trained on the prostate cancer dataset was used to label the obtained SNVs of the glioma cfDNA samples. The classifier predicted the type of SNVs with an AUC of 96 % (Fig. 2C) which highlights its adaptability and effectiveness across different types of cancer, thereby confirming the model's capability to generalize well beyond the initial training data. This high performance across distinct datasets is a testament to the model's robustness and the soundness of our methodological approach. As mentioned earlier, SSL aims to make use of large amounts of unlabeled data to boost model performance, typically when obtaining labeled data is expensive and time-consuming. Accordingly, various semi-supervised learning methods have been proposed using deep learning and have proven to be successful on several standard benchmarks [18]. In this paper, we focused on GAN-based SSL models and performed a comparison between our SSGAN model performance and other classification and semi-supervised classification methods such as MLPRegressor [19], BayesianRidge [20], and Kneighborsregressor [21]. In our study, we chose MLPRegressor, BayesianRidge, and KNeighborsRegressor as baseline models to compare with the performance of our proposed SSGAN model. These models were selected due to their distinct methodological approaches, each offering a unique perspective on handling the semi-supervised classification task. Below, we explain the importance of comparing these models in the context of our work.

2.3. *MLPRegressor (Multilayer Perceptron Regressor)*

- **Relevance to Genomic Data**: The MLPRegressor is a powerful feedforward neural network model that can capture complex, non-linear relationships between input features. In the context of genomic data, this is crucial because genomic interactions and patterns of somatic mutations are inherently complex and non-linear. The ability of MLPs to handle multiple layers of abstractions is a key factor in genomic studies, where hidden patterns might be obscured within layers of biological processes.
- **Comparison Rationale**: MLPRegressor allows us to benchmark our SSGAN model's performance against a highly flexible, non-linear model. While MLPs can excel in capturing complex relationships, they are also sensitive to overfitting, especially with small datasets. In contrast, our SSGAN model addresses this challenge by generating synthetic data to augment the dataset, which helps improve generalization. Therefore, the comparison between MLPRegressor and SSGAN helps demonstrate the effectiveness of our model in dealing with data scarcity and complex relationships, common in genomic datasets.

2.4. *BayesianRidge*

- **Relevance to Genomic Data**: BayesianRidge is a linear regression model that incorporates Bayesian principles to estimate the distribution of model parameters. In genomics, where datasets are often small and prone to overfitting, Bayesian methods provide a robust alternative by incorporating prior knowledge and providing uncertainty estimates for predictions. This model is particularly useful in semi-supervised learning scenarios, as it can effectively manage small labeled datasets by incorporating prior distributions.
- **Comparison Rationale**: We included BayesianRidge to compare our SSGAN's ability to model complex genomic relationships with a more traditional, probabilistic model that is robust in the face of data scarcity. While BayesianRidge works well for linear relationships and small datasets, its limitations in handling complex, non-linear genomic interactions make it a useful but insufficient model for our application. The comparison highlights how our SSGAN model can outperform simpler probabilistic approaches when dealing with large-scale genomic data that requires capturing more complex dependencies between variables.

2.5. *KNeighborsRegressor*

- **Relevance to Genomic Data**: KNeighborsRegressor is a non-parametric model that makes predictions based on the proximity of data points in feature space. In the context of genomic data, where different mutation types might cluster together, this model can be effective for simple, localized patterns. It has the advantage of being straightforward to implement and interpret, particularly in cases where the data has clear boundaries or groupings (e.g., specific mutation clusters).

- **Comparison Rationale**: KNeighborsRegressor serves as a benchmark for our SSGAN by providing a direct, distance-based classification approach. However, its limitations in high-dimensional datasets, like genomic data, where the curse of dimensionality can significantly degrade performance, highlight the need for more sophisticated models like SSGAN. By comparing with KNeighborsRegressor, we aim to showcase how our model overcomes the challenges posed by large, high-dimensional genomic datasets, providing more accurate and scalable solutions for mutation classification.

Based on the above-mentioned descriptions, our **SSGAN model** excels by combining semi-supervised learning with generative modeling, allowing it to generate synthetic data and learn from huge amount of unlabeled data. This approach significantly enhances performance in genomic classification tasks, particularly when labeled data is scarce. The comparisons emphasize SSGAN's ability to generalize well across different contexts, a critical requirement in the field of genomics, where mutation patterns and biological processes are often highly complex and not easily captured by simpler models. Ultimately, these comparisons help to establish the robustness of our SSGAN model, showcasing its superiority in capturing the intricate patterns of genomic mutations, especially in semi-supervised scenarios where data labeling is resource-intensive and time-consuming. The accuracy metric of all the methods increased as the input size enriched however, the methods implemented based on neural networks like SSGAN and MLPRegressor showed better performances. This is likely due to the ability of neural networks to capture complex patterns and relationships in the data. As illustrated in Fig. 2D, by integrating GANs and SSL, we were able to achieve better results with higher accuracy around 95 % in test phase of model development.

Next, we set out to further validate the performance of our classifier by comparing to matched tumor and WBC samples. Initially, our model was developed using a dataset related to pancreatic cancer. To validate the generalizability and robustness of our SSGAN model, we subsequently tested it on a distinct and independent dataset related to glioma cancer. As illustrated in Fig. 3, the model produced a noticeable and consistent pattern across both the prostate and glioma cancer datasets. Specifically, in both cases, there is a distinct separation between tumor-labeled and CH-labeled variants when compared against the real matched tumor and WBC samples. The overlapping fraction of detected variants is higher for tumor-labeled variants in real tumor samples and higher for CH-labeled variants in real WBC samples. This consistent behavior across different cancer types supports the reliability of our model in distinguishing between CH and tumor-derived mutations in cfDNA, regardless of the specific cancer type. This additional validation using a glioma cancer dataset not only reinforces the pattern observed but also demonstrates the versatility of our SSGAN model across different types of cancer. By doing so, we increase confidence in the model's ability to generalize beyond the initial dataset and its potential utility in various clinical contexts. For this analysis, the output of SSGAN model containing CH or tumor-labeled SNVs for cfDNA samples were divided into two distinct files and were compared against their real matched tumor and WBC samples. As illustrated in Fig. 3A, the tumor-labeled SNVs of prostate cancer dataset by SSGAN were mostly detected in the real tumor samples and the opposite was observed in CH-labeled SNVs which show the most concordance with real WBC samples. The same pattern was also detected when running the SSGAN model to classify the SNVs of glioma cfDNA dataset as mentioned earlier (Fig. 3B).

In the next step, all the mentioned variants were gone under mutation signature analysis due to its importance in cancer genomics. De novo mutation signature analysis by Mutalisk [22] was applied for each set of real and predicted SNVs. The difference between the mutation signatures distributions in CH and tumor-predicted SNVs for both prostate and glioma cancer dataset showed a positive correlation with their matched tumor and WBCs (with a Pearson correlation coefficient around 0.7853) (Supplementary Material 3, Figure A and B). Due to the obtained results, it can be inferred that there are no significant differences between the mutation signature patterns in different SNV types (mean adjusted P-value around 0.8) however the most meaningful difference was observed between CH-labeled and real tumor-derived variants (Supplementary Material 3, Figure C).

The classified SNVs by SSGAN were also annotated with ANNOVAR [23] and the obtained gene lists were assessed. A list of previously reported CH-driver genes including 64 genes was obtained from IntOGen (https://intogen.org/ch/search). 48 % and 68 % of these CH reference genes were detected in the cfDNA and WBCs samples of prostate cancer dataset, respectively (Fig. 4A), and 38 % of them have been observed in CH-labeled SNVs with 7 % in matched tumor-labeled variants. Meanwhile, in the glioma dataset, the

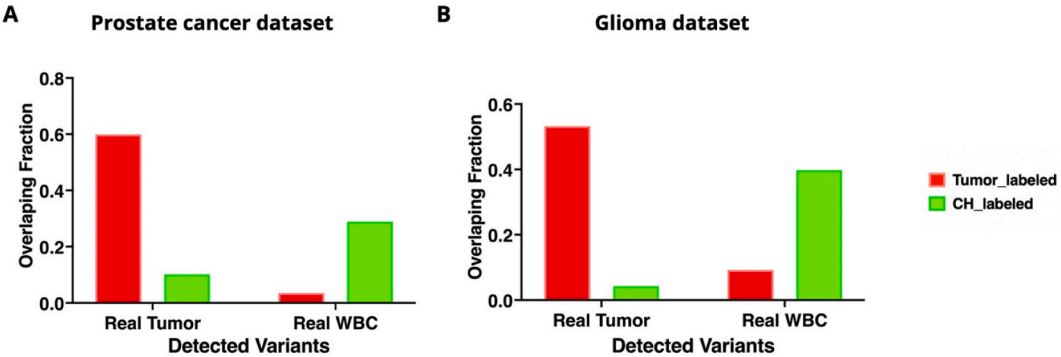


Fig. 3. Investigation of the overlap between labeled cfDNA SNVs by SSGAN model and real data. **A)** The grouped bar plot illustrating the proportion of overlap between SSGAN labeled SNVs and real tumor, real WBC in prostate cancer dataset. **B)** The grouped bar plot illustrating the proportion of overlap between SSGAN labeled SNVs and real tumor, real WBC in glioma dataset.

concordance between the Intogen database and CH-labeled genes expanded to encompass 36 genes, constituting 56 % of the reference gene list. It is worth highlighting those two pivotal genes, DNMT3A and TET2, each harboring several loci associated with CH [9] were accurately identified as CH-derived by SSGAN. To comprehensively evaluate the performance of our model, we compiled lists of control genes for each mutation type. First, a random gene set from previously documented housekeeping genes [24] was collected. A random mixture of 64 genes was selected from the aforementioned compilation of human housekeeping genes, and their presence in real WBC and CH-labeled data was examined (Fig. 4B). Next, a prostate cancer driver gene list was acquired from cBioPortal prostate adenocarcinoma TCGA PanCancer data with 745 distinct genes. Then, we investigated the state of this gene panel in real tumor and tumor-labeled variants (Fig. 4C). Meanwhile, a glioma gene list was obtained from cBioPortal from brain lower grade glioma (TCGA, PanCancer Atlas and the highly recommended oncogenes with a frequency of more than 50 % were extracted from the data and their status was investigated in tumor-labeled genes from our model. By applying these filtrations, the number of genes in the glioma gene panel reduced to 250. Among the 672 tumor-labeled genes and 250 glioma genes, 107 were detected as common genes with SDHA, TEC, MTOR, PIK3CD, FANCD2, CAMTA1, SPEN were more frequent. It is essential to underscore that these genes are associated with tumor suppression and the regulation of cell proliferation processes [25–31]. Next, a selection of less probable oncogenes was curated from cBioPortal, each boasting a mere 0.2 % frequency, was designated as the negative control for tumor-labeled mutations. This assemblage of negative controls comprised 5270 genes, of which only 159 were detected among the tumor-labeled mutations, each featuring a frequency of less than 0.5 % in our dataset (Fig. 4D).

Next, we performed gene set enrichment analysis on both the real and the SSGAN-labeled SNVs by using g: Profiler [32] and biological pathway enrichment was conducted for annotated variants. In a recent study using gene expression data, a positive correlation between CH status and increased level of inflammation factors was observed, mostly related to TET2 or DNMT3A mutations [14]. In our prostate cancer dataset, 'chemotaxis' and 'cellular response to stimulus' were also enriched in CH-labeled SNVs by our model which is consistent with previous findings (Supplementary Material 4, Figure A and B). The biological pathways enriched in the tumor-derived variants also showed meaningful results related to cell regulation and amplifications (Supplementary Material 4, Figure C and D).

2.6. Simulating new cfDNA variants using SSGAN

In addition to developing a classifier for predicting somatic variant types, our SSGAN model is able to simulate new cfDNA SNVs as well. NGS data simulators play a pivotal role in advancing genomics research and bioinformatics tool development. Serving as

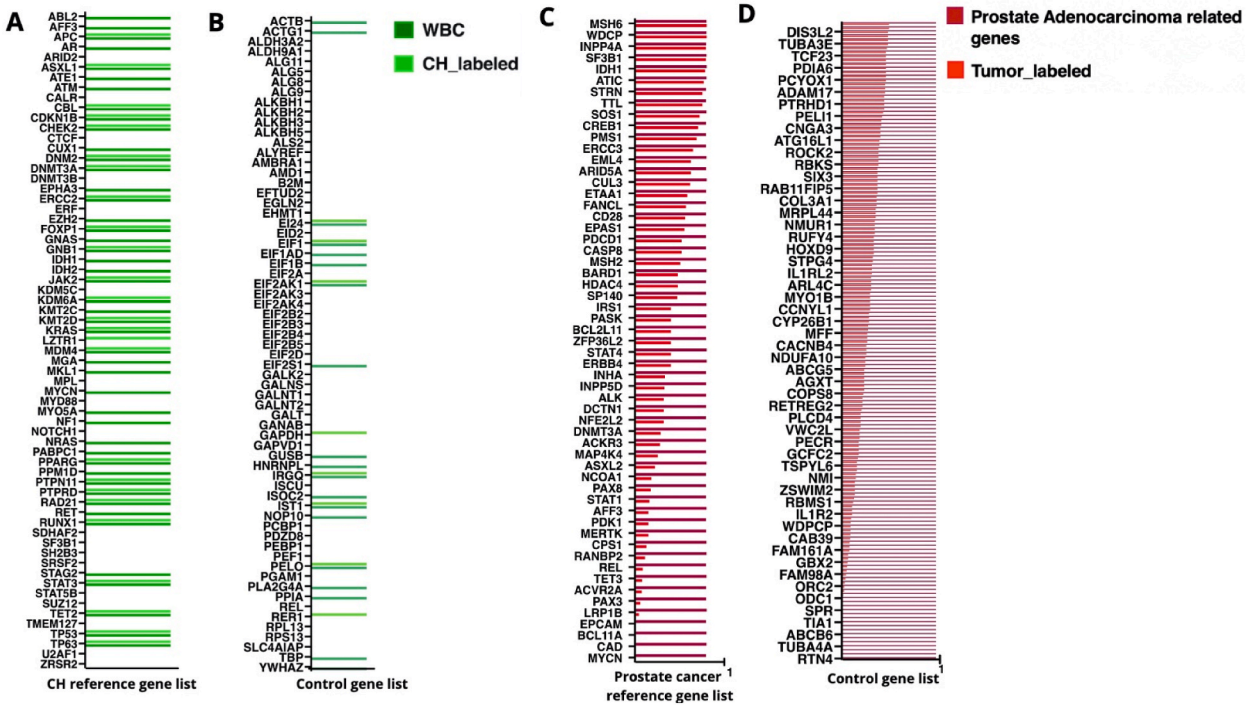


Fig. 4. Gene annotation and investigation analysis of labeled cfDNA SNVs by SSGAN model. **A)** The plot showing the presence of each IntOGen-reported CH-related gene in real WBC and CH-labeled variants by SSGAN model in prostate cancer dataset. **B)** The plot showing the presence of each human housekeeping-reported gene as a control group for validated CH-related reference genes in real WBC and CH-labeled variants by SSGAN model in prostate cancer dataset. **C)** The bar plot showing the frequency of each TCGA-reported prostate cancer-related gene in real tumor and tumor-labeled variants by SSGAN model in prostate cancer dataset. **D)** The bar plot showing the frequency of each cBioportal-reported less-probable oncogenes and tumor-labeled variants by SSGAN model in prostate cancer dataset.

controlled environments, these simulators generate synthetic datasets mirroring the characteristics of real NGS data. The significance of NGS data simulators lies in their ability to benchmark and optimize bioinformatics tools, validate new methods, and assess sensitivity and specificity under various conditions. The simulated datasets can be used to study the impact of experimental design factors, such as sequencing depth and read length, on analysis pipelines. In this study, the GAN engine of SSGAN can be viewed as a simulator for other research goals like benchmarking novel variant callers and variant annotation tools. Here, the simulated cfDNA variants were compared to real cfDNA data based on different genomic and functional characteristics. The loss value of model which is calculated based on cross-entropy was decreased by each epoch which shows the convergence of the generator and discriminator (Supplementary Material 5, Figure A). The model loss which is related to the GAN performance was measured both for prostate cancer (training: 0.22, test: 0.286) and glioma (training: 0.176, test: 0.210) dataset separately. Along with the model computational metrics, we have investigated the content of the generated SNVs as well. For this aim, first the chromosome type distribution was extracted from both real and simulated cfDNA SNVs and compared with each other as boxplots in Supplementary Material 5, Figure B which leads to no significant difference ($P > 0.05$). Then, the landscape of nucleotide conversion was compared between the simulated and real cfDNA variants and both showed no significant difference ($P > 0.05$) and high concordance (correlation coefficient around 0.9435) when considering the same sample size (Supplementary Material 5, Figure C).

3. Discussion

Liquid biopsy, an emerging experimental approach for detecting somatic mutations, is capable of serving as a valuable tool for cancer diagnosis and monitoring through the identification of genomic and epigenomic biomarkers. However, the presence of CH mutations in plasma can significantly impact the ability to accurately identify true positive somatic mutations in the area of cancer genomics. Consequently, there is a need for experimental or computational methods capable of discriminating plasma somatic mutations from each other based on their tissue of origin, presenting an area of high interest and application in cancer research and diagnostics. The cost associated with identifying plasma mutation types is a significant barrier due to the requirement for multiple distinct samples of an individual. Machine learning models present a solution, offering the means to overcome these obstacles. Machine learning's capacity to process large volumes of data efficiently and accurately positions it as a powerful tool in situations where labeling real data is resource-intensive and time-consuming. GANs have influenced the field of SSL by providing an effective means to extract valuable insights from limited input data. By integrating GANs into the SSL framework, we achieved greater accuracy and performance than traditional classification algorithms.

Our research centered around the development of a novel in silico approach for the classification of somatic variants found in cfDNA. We aimed to label these variants and also identify driver gene mutations, a crucial aspect of cancer research. A key outcome of our study was confirming our model's robustness and reproducibility. This was exemplified when we applied the model to diverse datasets representing different cancer types, further highlighting the consistency and reliability of the SSGAN framework. Our in-depth annotation of the CH-labeled SNVs led to the revelation of their involvement in essential biological processes. These CH-labeled SNVs were linked to chemotaxis and responses to stimuli. This discovery aligns with recent research findings [14] that highlight an increased level of inflammation markers associated with CH mutations. Such biological insights exemplify the broader implications of our research and its potential impact on our understanding of CH and its role in health and disease. Our study emphasizes the broader perspective that somatic variants are not confined to cancer samples but extend to healthy tissues as well. SomaMutDB is a resource for categorizing somatic mutations in various healthy tissue types [33]. In this regard, we propose expanding our predictive model to encompass the multi-classification of cfDNA somatic variants. This expansion involves integrating information from SomaMutDB with our established SNV catalog, promising to further advance the field. Meanwhile, by accurately distinguishing between CH and tumor-derived mutations, SSGAN model holds significant potential to reduce the incidence of false positives in liquid biopsy applications. This enhancement directly contributes to improving the precision of early cancer detection, which is crucial for making informed clinical decisions, leading to more accurate diagnoses and personalized treatment plans. The ability of our SSGAN model to differentiate between these mutation types in cfDNA could profoundly impact patient management, particularly in monitoring disease progression and assessing response to therapy. For instance, accurate detection of tumor-derived mutations can lead to earlier intervention, potentially improving patient prognosis. Conversely, reducing the misclassification of CH mutations as tumor-derived can prevent unnecessary treatments, thereby minimizing patient risk and healthcare costs. To facilitate the clinical adoption of SSGAN model, we outline a comprehensive pathway that includes several key steps. First, extensive validation across diverse and more independent cfDNA datasets is necessary to ensure the model's robustness, generalizability, and reproducibility in real-world clinical settings. This would involve collaboration with clinical laboratories and research institutions to access varied datasets and test the model's performance in different patient populations and cancer types. Second, integration into existing diagnostic workflows would require careful consideration of compatibility with current liquid biopsy protocols and infrastructure. This may involve adapting the model to work seamlessly with widely-used cfDNA sequencing technologies and ensuring it meets regulatory requirements for clinical use. Additionally, the development of user-friendly software tools that incorporate our SSGAN model could facilitate its adoption by clinicians, making it easier to interpret and act on the results. Finally, by providing a more accurate method for interpreting cfDNA data, our model can contribute to more timely and appropriate clinical interventions. For instance, the SSGAN model could be used in conjunction with existing diagnostic tests to confirm the presence of cancer-related mutations, or to monitor minimal residual disease (MRD) in patients undergoing treatment. Over time, as the model is refined and its predictive accuracy further validated, it could become an integral component of personalized cancer care, supporting precision oncology initiatives that tailor treatments based on an individual's unique genetic profile. In conclusion, our study underscores the transformative potential of machine learning in the interpretation of complex genomic data, particularly in scenarios where labeling real data is resource-intensive and time-consuming.

Another advantage of the SSGAN model lies in its ability to generate novel cfDNA SNVs, serving as an independent dataset that mimics the characteristics of real cfDNA variants.

4. Conclusion

In this study, we introduced a deep learning model called SSGAN to enhance the precision of somatic variant classification in cfDNA samples. Our model addresses a critical gap in liquid biopsy analysis by distinguishing CH from tumor-derived SNVs, a distinction that has posed significant challenges and resource demands in traditional methodologies. The SSGAN model's adaptability was demonstrated through performance assessment across multiple cancer datasets, confirming its reliability and accuracy. Additionally, our approach contributes to the bioinformatics field by generating simulated cfDNA SNVs, offering a novel resource for benchmarking variant analysis tools in the area of cancer cfDNA analysis.

4.1. Materials and methods

In this study, we tried to provide a computational method to identify the origin of cfDNA somatic variants (CH or tumor) instead of cost vector laboratory methods. For this purpose, machine learning algorithms can be a good choice that can help us to achieve this goal by providing appropriate input data and implementing a good computational architecture. In this regard, at first, we collected the proven labeled variants from studies and other databases and create a somatic SNV catalog, but since their number was limited, the semi-supervised algorithm was proposed to carry out the variant classification task, and is explained in more detail below.

4.2. Data preparation

Due to the semi-supervised nature of our model, we have two main groups of data points. One is the unlabeled group which includes the somatic variants obtained from cfDNA WGS samples and the labeled group which is created based on the somatic variants that are experimentally validated to be derived from CH or tumor. The first group included genetic material derived from cfDNA, tissue, and WBC samples, all originating from the same individuals and the second part included the variants that were validated through experimental techniques and were categorized as either CH or tumor-derived variants. To prepare the unlabeled part of the proposed model, we accessed raw FASTQ files from two distinct sources: the European Genome-phenome Archive (EGA) with accession number EGAD00001004526 [16] for prostate cancer and the other one from National Center for Biotechnology Information (NCBI) with accession number SRP268702 for glioma. The prostate cancer dataset served as the foundation for developing our model, while the glioma cancer dataset was reserved for subsequent model validation. On the other hand, we tried to create the labeled dataset for the classification task of our model. For this aim, we have collected the catalog of somatic variants as ground truth data that each of the variants is classified as CH or tumor. For the tumor somatic class, we compiled a diverse set of cancer-related variants from reputable sources such as cBioPortal and the Catalogue Of Somatic Mutations In Cancer (COSMIC). In the case of the CH class, we established our reference dataset by drawing upon validated results from multiple studies, as indicated in Supplementary Material 1 [7, The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020, 9]. This locally-curated somatic variant catalog then underwent the redundancy removal and the overlapped variants between the two classes (tumor and CH) were removed which finally leads to of 5000 non-redundant variants. The specificity of the variants for each class is on the mutation level and not gene level. As one gene can harbor completely different variants related to tumor or CH origin.

4.3. Variant calling of genomic data

We conducted somatic variant calling on the raw FASTQ sequencing reads following the Genome Analysis Tool Kit (GATK) (McKenna et al., 2010) best practices [34] by using GRCH38 reference genome. The primary variant calling algorithm used in this process was Mutect2[35]. The resulting variant calling format (VCF) files were subjected to a particular filtering procedure based on three fundamental criteria.

1) SNV: variants had to meet the criterion of being only SNV for inclusion in the subsequent analysis and indels were not considered in this study.
2) Human chromosomes: only variants associated with the 23 pairs of chromosomes and the mitochondrial genome were retained and other scaffolds in the reference genome were ignored.
3) High depth: variants were further filtered to ensure they exhibited a high depth, with a depth exceeding 30.

We used the 20 VCF files of cfDNA prostate cancer dataset that passed these stringent filters for our model construction. The other VCF files, representing tissue or WBCs were reserved for evaluating the model's performance and biological interpretation of the results. Following GATK's best practices for somatic variant calling [36], we identified approximately 90,000 variants in each prostate cancer cfDNA sample. Following the application of our filtration steps, this number was refined to around 70,000 SNVs for subsequent analysis and interpretation.

4.4. Selecting genomic features

After preprocessing the raw cfDNA sequencing samples and obtaining the SNVs, we needed to prepare the feature set of our model based on genomic characteristics. In this study, we aimed to build a model based on simple features that can be obtained by straightforward processing methods and not make the feature set so complex by adding other information such as long nucleotide sequences of gene annotations which requires more processing before and for model running. Accordingly, we integrated the variant's genomic coordination along with nucleotide type information. We introduced a new feature as the nucleotide composition around the single variant position. To obtain this nucleotide composition, we utilized the GRCH38 reference genome, samtools faidx [37], and a few Linux commands to extract sequences based on the variant's coordinates. We conducted experiments with various window sizes, including 10, five, and three bases upstream and downstream of each variant's position. The model demonstrated its optimal performance when a window size of three bases was incorporated into the feature space. This approach resulted in assigning a 10-feature vector (chromosome number, the SNV position, reference nucleotide at the SNV position, alternate nucleotide at the SNV position, three nucleotides upstream of the SNV position, and three nucleotides downstream of the SNV position) to each data point.

4.5. Implementing the SSGAN model

Given the insufficiency of labeled data in comparison to the substantial number of unlabeled variants constituting only 10 % of the total dataset, conventional supervised algorithms were not considered to be suitable for the binary classification task in the current study. Consequently, we adopted semi-supervised learning (SSL) methods [38] as a suitable strategy for cfDNA somatic variant classification. While there are several SSL algorithms available, recent approaches that integrate SSL with generative models have exhibited superior performance [17]. This is largely due to their ability to encompass a representation learning phase in conjunction with classification accordingly the model can learn different and hidden aspects of the dataset and apply the classification task in the following step more accurately. In our study, we found inspiration in semi-supervised learning with GAN architecture [15] and customized it to align with our classification task. The algorithm of our implemented semi-supervised generative adversarial networks (SSGAN) model is depicted in Algorithm 1. The following is a detailed breakdown with specific parameter values for Algorithm 1.

1. Initialization:
   - The algorithm begins by initializing the parameters for both the generator and discriminator networks within the SSGAN framework.
   - **Learning Rate**: The learning rate for both the generator and discriminator is set to $\eta = 0.05$.
   - **Optimizer**: The Adam optimizer is used.
2. Input Data Preparation:
   - **Labeled Data**: The labeled dataset consists of **30 %** of the input data, where each sample is categorized as either CH-related or tumor-derived and only fed to the classifier.
   - **Unlabeled Data**: The remaining **70 %** of the input data is unlabeled, serving as the basis for semi-supervised learning and is the input for the discriminator.
   - This mix of labeled and unlabeled data is crucial for enabling the model to generalize effectively and improve classification performance.
3. Generator Training:
   - The generator network is initialized to create synthetic data samples that closely resemble the real cfDNA variants. The generator receives random noise vectors following a normal distribution as input and produces data mimicking cfDNA somatic mutations.
   - **Batch Size**: The model is trained using a batch size of **2000 samples** per iteration.
4. Discriminator Training:
   - The discriminator network is trained concurrently with the generator. It receives both real and synthetic cfDNA samples as input and aims to distinguish between them.
   - The discriminator's loss function includes two components: one for classifying real vs. synthetic data and another for correctly classifying the labeled real data in the classification phase and tuned up for the GAN part of the model.
   - The discriminator's training is structured to balance its ability to classify and differentiate between the real and synthetic data effectively.
5. Semi-Supervised Learning Process:
   - After training the GAN, the discriminator tuned architecture now is used to classifies the labeled data with accuracy.
   - This process allows the model to improve its classification ability with limited labeled data, exploiting the large amounts of unlabeled data available.
   - **Epochs**: The training process runs for **1000 epochs** to ensure sufficient learning and convergence of the model.
6. Loss Calculation and Optimization:
   - The loss functions for both the generator and the discriminator are computed during each iteration based on cross entropy.
   - **Discriminator Loss**: This includes a component for classifying real vs. synthetic data and another for classification accuracy on only labeled data.
   - **Generator Loss**: This loss is focused on the discriminator's ability to distinguish real from synthetic data.
   - The Adam optimizer is used to iteratively update the model parameters to minimize these losses, facilitating convergence.

7. Convergence and Model Outputs:
   - o Training continues until the model reaches convergence, meaning the generator produces highly realistic cfDNA samples that the discriminator struggles to distinguish from real data.
   - o The final SSGAN model is capable of accurately classifying somatic mutations in cfDNA, distinguishing between CH-related and tumor-derived mutations.

In this study, the SSGAN model was implemented by TensorFlow library, the GAN part of which was aimed at identifying the hidden parts of the data and applying the representation learning of cfDNA SNVs. While the classification part was for classifying cfDNA somatic SNVs as CH or tumor. In this way, the feature matrix including only unlabeled somatic SNVs of prostate cancer with a dimension of 10 x 100,000, first produced based on the preprocessing method mentioned above and normalized for position values then given to the discriminator engine of the GAN, which tried to label these data points as real or simulated by comparing with the one initially generated by the generator based on random samples from the normal distribution. It is worth mentioning that unlabeled data provides additional information that helps the model to learn more robust and discriminative features. In the context of the SSGAN, the unlabeled somatic SNVs of prostate cancer serve as valuable examples for the model to learn the underlying patterns and representations within the data. By exposing the model to a wider range of data points, it can extract more meaningful features, leading to better generalization and improved performance. The architecture used in the generator is 9 layers deep neural networks and for discriminator is six convolutional layers that are designed to extract hierarchical features from the input data. Both the generator and discriminator were trained with LeakyReLU activation function.

Initially, the generator and discriminator components undergo a training phase. Once they reach an optimal level of performance, then the pre-built discriminator engine is deployed for executing the binary classification task as well. All three constituent parts of our model are constructed on deep neural networks and Adam optimizer [39] along with learning rate $\eta = 0.05$ were implemented in classification task. Here, the loss function of each deep neural network including the loss function of the generator, discriminator, and classifier are provided in Equations (1)–(3).

Loss function of generator:

$$L_G = -\frac{1}{m}\sum_{i=1}^{m} log\, D\left(G\left(z^{(i)}\right)\right) \tag{1}$$

In this formula, m denotes the number of generated data points (SNVs) by the generator based on a normal distribution, and z is a matrix related to a random sample from the hidden space (generated samples). G and D indicate generator and discriminator, respectively.

Loss function of discriminator:

$$L_D = -\frac{1}{m}\sum_{i=1}^{m} log\left(1 - D\left(G\left(z^{(i)}\right)\right)\right) - \frac{1}{n}\sum_{i=1}^{n} log\, D\left(x^{(i)}\right) \tag{2}$$

Here, m is again related to the number of generated data points by the generator, and n denotes the number of real SNVs from the unlabeled part of the dataset that are the input of the discriminator and obtained from the cfDNA WGS samples. x is a matrix related to the real sample from the unlabeled part of the dataset.

Loss function of classifier:

$$L_C = -\frac{1}{p}\sum_{i=1}^{p} \left(y^{(i)}\, log\, C\left(x^{(i)}\right) + \left(1 - y^{(i)}\right) log\, \left(1 - C\left(x^{(i)}\right)\right)\right) \tag{3}$$

In this formulation, p denotes the number of labeled SNVs that are obtained from our locally developed somatic variant catalog. C showes the classifier which has the same architecture of discriminator and use the real labeled SNVs with label y to carry out the cfDNA SNV binary classification task.

Accordingly, the model receives the cfDNA SNV information as a matrix including the previously described 10 features for each data point and which provides two different outputs. One output is the simulated cfDNA SNVs created by the GAN and the other is the input cfDNA SNVs with the label of their class as tumor or CH.

Algorithm 1: SSGAN

for *each epoch* **do**
  Sample $z^1, ..., z^m$ from $\mathcal{N}(0, 1)$;
  Sample $x^1, ..., x^n$ from real data;
  $g_{w_G} \leftarrow \frac{1}{m}\nabla_{w_G}\sum_{i=1}^{m}[logD(G(z^i))]$;
  $w_G \leftarrow w_G + \eta.adam(g_{w_G}, w_G)$;
  $g_{w_D} \leftarrow \frac{1}{m}\nabla_{w_D}\sum_{i=1}^{m}[log(1 - D(G(z^i)))] + \frac{1}{n}\nabla_{w_D}\sum_{i=1}^{n}[logD(x^i)]$;
  $w_D \leftarrow w_D + \eta.adam(g_{w_D}, w_D)$;
end
for *each epoch* **do**
  Sample $(x^1, y^1), ..., (x^p, y^p)$ from Labeled data;
  $g_{w_C} \leftarrow \frac{1}{p}\sum_{i=1}^{p}[y^i logC(x^i) + (1 - y^i)log(1 - C(x^i))]$;
  $w_C \leftarrow w_C + \eta.adam(g_{w_C}, w_C)$;
end

4.6. Evaluation of SSGAN model performance

Our evaluation of the SSGAN model's performance followed a dual approach. Initially, we assessed its computational performance using conventional metrics derived from loss functions. Subsequently, we embarked on a biological interpretation of the results, which took into account information from matched tumor and WBC samples. From a computational perspective, we assessed the model's performance through well-established metrics, including accuracy, precision, and the area under the curve (AUC). Particularly, given the utilization of the GAN structure [40] in our study, we also computed metrics related to loss based on cross-entropy. Transitioning to the biological interpretation, we proceeded by defining the overlaps between the model's outputs and the variants identified in matched tumor and WBC samples. This involved the utilization of VCF tools [41] along with custom Python scripts available on our GitHub page. Furthermore, we carried out the functional assessment of the classified variants. For this purpose, we exploited tools such as Mutalisk [22] for mutational signature analysis and ANNOVAR [23] for somatic variant annotation. These analyses were applied to both real tumor samples and SNVs labeled by the SSGAN machine. Statistical analyses were then performed to assess the similarity between the two data sources by Prism. An analogous analysis pipeline was executed for CH-labeled SNVs compared to matched real WBC samples.

4.7. Validation with an independent dataset

As previously mentioned, the unlabeled part of our feature set was created from the prostate cancer cfDNA samples and the labeled part from the locally developed somatic SNVs catalog. In order to confirm the reliability and generalizability of our model, we extended our validation to encompass an independent dataset related to glioma, a form of brain cancer encompassing three cfDNA samples, each accompanied by corresponding matched tumor and WBC samples. This dataset underwent the same GATK workflow for somatic variant calling and following filtration steps. After the preprocessing phase, the obtained SNVs from glioma cfDNA samples were fed to the GAN part of the model as an unlabeled dataset. Another model validation was also done by implementing divers machine learning models for cfDNA SNV type classification however, given the distinct nature and features of the methods employed in our study, we focused solely on validating our model using distinct cfDNA samples because there was no other similar method for this kind of study.

CRediT authorship contribution statement

Fahimeh Palizban: Writing – review & editing, Visualization, Validation, Methodology, Conceptualization. **Mohammadmahdi Sarbishegi:** Methodology. **Kaveh Kavousi:** Supervision. **Mahya Mehrmohamadi:** Writing – review & editing, Supervision, Conceptualization.

Data availability

All processed data are available as Supplementary Materials. Raw sequencing data were obtained from EGAD00001004526 and SRP268702.

Code availability

The codes to replicate the findings in the article are publicly available on: https://github.com/FPalizban/SSGAN. Also, to run the model automatically the Google Colaboratory notebook of the SSGAN is available on: https://github.com/FPalizban/SSGAN/blob/main/SSGAN.ipynb.

Funding

This research received no specific grants from any funding agency.

Declaration of competing interest

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

Acknowledgments

The authors would like to appreciate the assistance of Kaveh Motamedian in designing the schematic illustration of SSGAN and Sina Majidian for his valuable comments and constructive feedback on this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e39379.

References

[1] S.N. Lone, S. Nisar, T. Masoodi, M. Singh, A. Rizwan, S. Hashem, M.A. Macha, Liquid biopsy: a step closer to transform diagnosis, prognosis and future of cancer treatments, Mol. Cancer 21 (1) (2022) 79.
[2] I.S. Haque, O. Elemento, Challenges in using ctDNA to achieve early detection of cancer, bioRxiv 237578 (2017).
[3] A.J. Bronkhorst, V. Ungerer, S. Holdenrieder, The emerging role of cell-free DNA as a molecular marker for cancer management, Biomolecular detection and quantification 17 (2019) 100087.
[4] S. Jaiswal, P. Fontanillas, J. Flannick, A. Manning, P.V. Grauman, B.G. Mar, B.L. Ebert, Age-related clonal hematopoiesis associated with adverse outcomes, N. Engl. J. Med. 371 (26) (2014) 2488–2498.
[5] A.G. Bick, J.S. Weinstock, S.K. Nandakumar, C.P. Fulco, E.L. Bao, S.M. Zekavat, P. Natarajan, Inherited causes of clonal haematopoiesis in 97,691 whole genomes, Nature 586 (7831) (2020) 763–768.
[6] J.E. Feusier, S. Arunachalam, T. Tashi, M.J. Baker, C. VanSant-Webb, A. Ferdig, C.C. Mason, Large-scale identification of clonal hematopoiesis and mutations recurrent in blood cancers, Blood cancer discovery 2 (3) (2021) 226–237.
[7] P. Razavi, B.T. Li, D.N. Brown, B. Jung, E. Hubbell, R. Shen, J.S. Reis-Filho, High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants, Nat. Med. 25 (12) (2019) 1928–1937.
[8] A. Leal, N.C.T. van Grieken, D.N. Palsgrove, J. Phallen, J.E. Medina, C. Hruban, M.A.M. Broeckaert, V. Anagnostou, V. Adleff, D.C. Bruhm, White blood cell and cell-free DNA analyses for detection of residual disease in gastric cancer, Nat. Commun. 11 (1) (2020) 525.
[9] S.P. Kar, P.M. Quiros, M. Gu, T. Jiang, J. Mitchell, R. Langdon, V. Iyer, C. Barcena, M.S. Vijayabaskar, M.A. Fabre, Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis, Nat. Genet. 54 (8) (2022) 1155–1166.
[10] O. Pich, I. Reyes-Salazar, A. Gonzalez-Perez, N. Lopez-Bigas, Discovering the drivers of clonal hematopoiesis, Nat. Commun. 13 (1) (2022) 4267.
[11] E. Xu, K. Su, Y. Zhou, L. Gong, Y. Xuan, M. Liao, J. Cao, Y. Li, Y. Lu, Y. Zhao, Comprehensive landscape and interference of clonal haematopoiesis mutations for liquid biopsy: a Chinese pan-cancer cohort, J. Cell Mol. Med. 25 (21) (2021) 10279–10290.
[12] K.L. Bolton, R.N. Ptashkin, T. Gao, L. Braunstein, S.M. Devlin, D. Kelly, M. Patel, A. Berthon, A. Syed, M. Yabe, Cancer therapy shapes the fitness landscape of clonal hematopoiesis, Nat. Genet. 52 (11) (2020) 1219–1226.
[13] A.S. Arun, J.E. Medina, S. Cristiano, D.C. Bruhm, R.J. Fijneman, G.A. Meijer, R.B. Scharpf, PLASMUT: an R Package for estimating the probability of tumor-specific mutations in cell-free DNA, Cancer Res. 84 (6_Supplement) (2024), 6101-6101.
[14] L. Fairchild, J. Whalen, K. D'Aco, J. Wu, C.B. Gustafson, N. Solovieff, O.A. Balbin, Clonal hematopoiesis detection in patients with cancer using cell-free DNA sequencing, Sci. Transl. Med. 15 (689) (2023) eabm8729.
[15] A. Odena, Semi-supervised Learning with Generative Adversarial Networks, 2016. ArXiv Preprint ArXiv:1606.01583.
[16] G. Vandekerkhove, W.J. Struss, M. Annala, H.M.L. Kallio, D. Khalaf, E.W. Warner, C. Herberts, E. Ritch, K. Beja, Y. Loktionova, Circulating tumor DNA abundance and potential utility in de novo metastatic prostate cancer, Eur. Urol. 75 (4) (2019) 667–675.
[17] Z. Dai, Z. Yang, F. Yang, W.W. Cohen, R.R. Salakhutdinov, Good semi-supervised learning that requires a bad gan, Adv. Neural Inf. Process. Syst. 30 (2017).
[18] X. Yang, Z. Song, I. King, Z. Xu, A survey on deep semi-supervised learning, IEEE Trans. Knowl. Data Eng. (2022).
[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, É. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[20] D.J.C. MacKay, Bayesian interpolation, Neural Comput. 4 (3) (1992) 415–447.
[21] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Statistician 46 (3) (1992) 175–185.
[22] J. Lee, A.J. Lee, J.-K. Lee, J. Park, Y. Kwon, S. Park, H. Chun, Y.S. Ju, D. Hong, Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures, Nucleic Acids Res. 46 (W1) (2018) W102–W108.
[23] K. Wang, M. Li, H. Hakonarson, ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data, Nucleic Acids Res. 38 (16) (2010), 164–e164.
[24] E. Eisenberg, E.Y. Levanon, Human housekeeping genes, revisited, Trends Genet. 29 (10) (2013) 569–574.
[25] N. Burnichon, J.J. Brière, R. Libé, L. Vescovo, J. Riviere, F. Tissier, A.P. Gimenez-Roqueplo, SDHA is a tumor suppressor gene causing paraganglioma, Hum. Mol. Genet. 19 (15) (2010) 3011–3020.
[26] L. Zeng, S. Yuan, J. Shen, M. Wu, L. Pan, X. Kong, Suppression of human breast cancer cells by tectorigenin through downregulation of matrix metalloproteinases and MAPK signaling in vitro, Mol. Med. Rep. 17 (3) (2018) 3935–3943.
[27] B.C. Grabiner, V. Nardi, K. Birsoy, R. Possemato, K. Shen, S. Sinha, D.M. Sabatini, A diverse array of cancer-associated MTOR mutations are hyperactivating and can predict rapamycin sensitivity, Cancer Discov. 4 (5) (2014) 554–563.
[28] J.S. Chen, J.Q. Huang, B. Luo, S.H. Dong, R.C. Wang, Z.K. Jiang, J.F. Zhong, PIK 3 CD induces cell growth and invasion by activating AKT/GSK-3β/β-catenin signaling in colorectal cancer, Cancer Sci. 110 (3) (2019) 997–1011.
[29] J. Moes-Sosnowska, I.K. Rzepecka, J. Chodzynska, A. Dansonka-Mieszkowska, L.M. Szafron, A. Balabas, J. Kupryjanczyk, Clinical importance of FANCD2, BRIP1, BRCA1, BRCA2 and FANCF expression in ovarian carcinomas, Cancer Biol. Ther. 20 (6) (2019) 843–854.

[30] K.O. Henrich, T. Bauer, J. Schulte, V. Ehemann, H. Deubzer, S. Gogolin, F. Westermann, CAMTA1, a 1p36 tumor suppressor candidate, inhibits growth and activates differentiation programs in neuroblastoma cells, Cancer Res. 71 (8) (2011) 3142–3151.
[31] S. Légaré, L. Cavallone, A. Mamo, C. Chabot, I. Sirois, A. Magliocco, M. Basik, The estrogen receptor cofactor SPEN functions as a tumor suppressor and candidate biomarker of drug responsiveness in hormone-dependent breast cancers, Cancer Res. 75 (20) (2015) 4351–4363.
[32] U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, J. Vilo, g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update), Nucleic Acids Res. 47 (W1) (2019) W191–W198.
[33] S. Sun, Y. Wang, A.Y. Maslov, X. Dong, J. Vijg, SomaMutDB: a database of somatic mutations in normal human tissues, Nucleic Acids Res. 50 (D1) (2022) D1100–D1108.
[34] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (9) (2010) 1297–1303.
[35] D. Benjamin, T. Sato, K. Cibulskis, G. Getz, C. Stewart, L. Lichtenstein, Calling somatic SNVs and indels with Mutect2, bioRxiv (2019) 861054.
[36] M.A. DePristo, E. Banks, R. Poplin, K.V. Garimella, J.R. Maguire, C. Hartl, A.A. Philippakis, G. Del Angel, M.A. Rivas, M. Hanna, A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (5) (2011) 491–498.
[37] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, Subgroup, 1000 Genome Project Data Processing, The sequence alignment/map format and SAMtools, Bioinformatics 25 (16) (2009) 2078–2079.
[38] O. Chapelle, B. Scholkopf, A. Zien, Semi-supervised learning, in: o. chapelle, et al. (Eds.), IEEE Trans. Neural Network. 20 (3) (2009) 542, 2006)[book reviews].
[39] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.
[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun. ACM 63 (11) (2020) 139–144.
[41] P. Danecek, A. Auton, G. Abecasis, C.A. Albers, E. Banks, M.A. DePristo, R.E. Handsaker, G. Lunter, G.T. Marth, S.T. Sherry, The variant call format and VCFtools, Bioinformatics 27 (15) (2011) 2156–2158.