


# A review of genetic variant databases and machine learning tools for predicting the pathogenicity of breast cancer

Rahaf M. Ahmad , Bassam R. Ali , Fatma Al-Jasmi , Richard O. Sinnott , Noura Al Dhaheri  and Mohd Saberi Mohamad 

Corresponding author: Mohd Saberi Mohamad, Tel.: +971 3 713 7654, Mob.: +971 50 129 0158, Fax: +971 3 7672 001. E-mail: [saberi@uaeu.ac.ae](mailto:saberi@uaeu.ac.ae)

## Abstract

Studies continue to uncover contributing risk factors for breast cancer (BC) development including genetic variants. Advances in machine learning and big data generated from genetic sequencing can now be used for predicting BC pathogenicity. However, it is unclear which tool developed for pathogenicity prediction is most suited for predicting the impact and pathogenicity of variant effects. A significant challenge is to determine the most suitable data source for each tool since different tools can yield different prediction results with different data inputs. To this end, this work reviews genetic variant databases and tools used specifically for the prediction of BC pathogenicity. We provide a description of existing genetic variants databases and, where appropriate, the diseases for which they have been established. Through example, we illustrate how they can be used for prediction of BC pathogenicity and discuss their associated advantages and disadvantages. We conclude that the tools that are specialized by training on multiple diverse datasets from different databases for the same disease have enhanced accuracy and specificity and are thereby more helpful to the clinicians in predicting and diagnosing BC as early as possible.

**Keywords:** pathogenicity prediction; genetic variants database; machine learning; artificial intelligence; breast cancer; data science

## INTRODUCTION

Machine learning (ML) is a subset of artificial intelligence that uses input data to learn patterns through many widely available algorithms and models. The challenges for analyzing and interpreting ever increasing volumes of data (big data) are increasing. Consequently, there is a need for novel ML tools to optimally process and learn from such big data. One emerging ML approach that is currently receiving much attention is deep learning (DL) [1]. It describes a family of algorithms/models, typically including multi-layer neural networks with many hidden units [2]. Such

models can be used to learn complex patterns that can, for example, support predictions [2].

Advances in technology, have changed the understanding of the available sequenced human genetic variants. Since the first human genome was sequenced, many more have been sequenced in academic, clinical and the private sector settings [3]. The number of rare variants is also growing and there is a pressing need to determine whether variants are pathogenic or benign.

In this context, breast cancer (BC) is one of the most common tumor types in the world [4]. In women between 20 and 50 years old, BC represents around 11% of all cancer mortalities [5], while

**Rahaf M. Ahmad** is a dedicated Ph.D. candidate in the Health and Data Science Lab at the United Arab Emirates University-College of Medicine and Health Sciences, she has a Bachelor Degree in Pharmacy, and a Master Degree in Nanoscience and Nanoengineering. With a diverse academic background, she brings a unique perspective to her pioneering research in breast cancer pathogenicity prediction using machine learning. Currently, she is in the 2nd year of her doctoral studies and she has seamlessly integrated her diverse background into her research. Her ability to navigate the interface multidisciplinary fields showcases her commitment to pushing the boundaries of research and innovation. Her academic journey is punctuated by publications, conference presentations and other research events. Her collaborative spirit is evident in her active participation in collaborative projects.

**Bassam R. Ali** is Professor of Molecular and Genetic Medicine and leader of the Genetics and Development Research Priority Group at the College of Medicine and Health Sciences, UAE University. Prof. Ali obtained his PhD degree in biochemistry from the University of Cambridge where he was the recipient of Said Foundation Scholarship and the UK ORS Award. Subsequently, Prof. Ali worked at University College London and Imperial College London before joining the UAEU in August 2006.

**Fatma Al Jasmi** is the Chair of the Genetic & Genomic department at UAE University and a Metabolic consultant at Tawam Hospital. She graduated from UAE University in 2000, pursued postgraduate studies at the University of Toronto, and received the Canadian Board of Pediatrics in 2006. As a founder of the UAE Rare Disease Society, she established the Biochemical Genetic Fellowship and has received numerous awards, including the Prime Minister Award for Excellence (2017) and the Women in Science Hall of Fame recognition (2015).

**Richard O. Sinnott** is the Director of eResearch at the University of Melbourne and Professor of Applied Computing Systems. In these roles he is responsible for all aspects of eResearch (research-oriented IT development) at the University. He has been lead software engineer/architect on an extensive portfolio of national and international projects, with specific focus on those research domains requiring finer-grained access control (security).

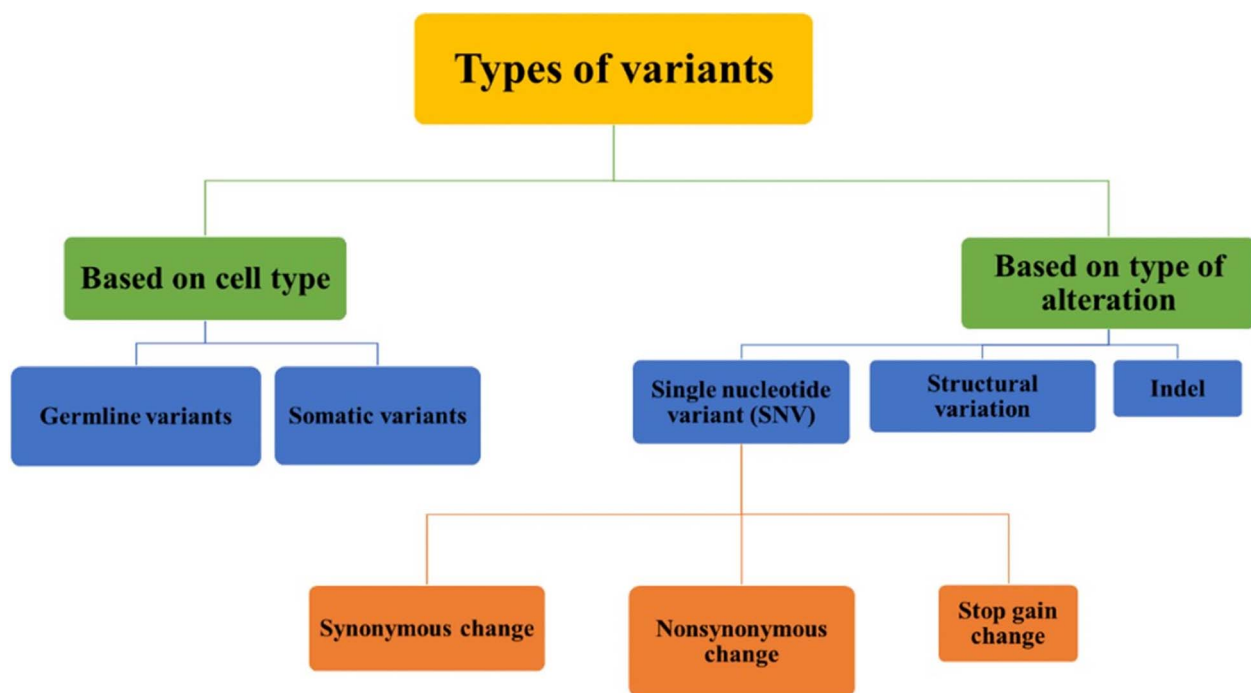
**Noura Al Dhaheri** is an Assistant Professor at the Department of Pediatrics, College of Medicine and Health Sciences, certified by the American Board of Pediatrics with qualifications in Medical Genetics & Genomics and Medical Biochemical Genetics. Her current research focuses on Genomic Medicine, particularly in novel gene discoveries for conditions like cleft lip/palate and vertebral malformation. Dr. Al Dhaheri is a part of the Baylor-Hopkins Center for Mendelian Genomics, aiming to elucidate the genetic basis of Mendelian disorders. Additionally, she explores Metabolomics to identify biomarkers for diagnosis and monitoring in patients with metabolic disorders, integrating this with genomics data for better variant annotation.

**Mohd Saberi Mohamad** is a professor of artificial intelligence and health data science. He is now the Director of the Health Data Science Lab in the College of Medicine and Health Sciences, UAE University. His research interests include artificial intelligence, bioinformatics and data science.

Received: June 22, 2023. Revised: September 22, 2023. Accepted: December 4, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1.** The types of variants based on cell type and alteration type.

in men, it is 19% higher compared with women [6]. The early diagnosis of BC to reduce the mortality rate is essential. In the field of medical analysis, ML algorithms have been extensively applied [7] with examples in predicting coronavirus disease (COVID)-19 [8], Alzheimer's progression [9], chronic diseases [9], liver disorders [10], heart disease [11], cancer [12] and others [13, 14]. The use of DL and ML for BC prediction is constantly advancing. The key factor in developing ML tools for BC lies in training them with specific BC data, rather than the algorithms themselves. Choosing the right ML tool for BC prediction is challenging due to the variability in datasets, which can impact the performance of ML models based on the training data [15]. A number of studies have explored ML prediction techniques for BC; however, these have not considered the pathogenicity of gene variants.

Human genetic variant databases serve as repositories of extensive data concerning thousands of human genetic variants, encompassing diverse information and purposes, from disease prediction [16] to supporting personalized medicine [17]. These databases, such as 1000 Genomes [18], COSMIC [19], ClinVar [20] and SwissVar [21] not only share variant-associated data but also maintain their unique annotations and datasets, resulting in heterogeneity across them. This diversity poses challenges in terms of data structure and consistency for geneticists, biologists and clinicians [22].

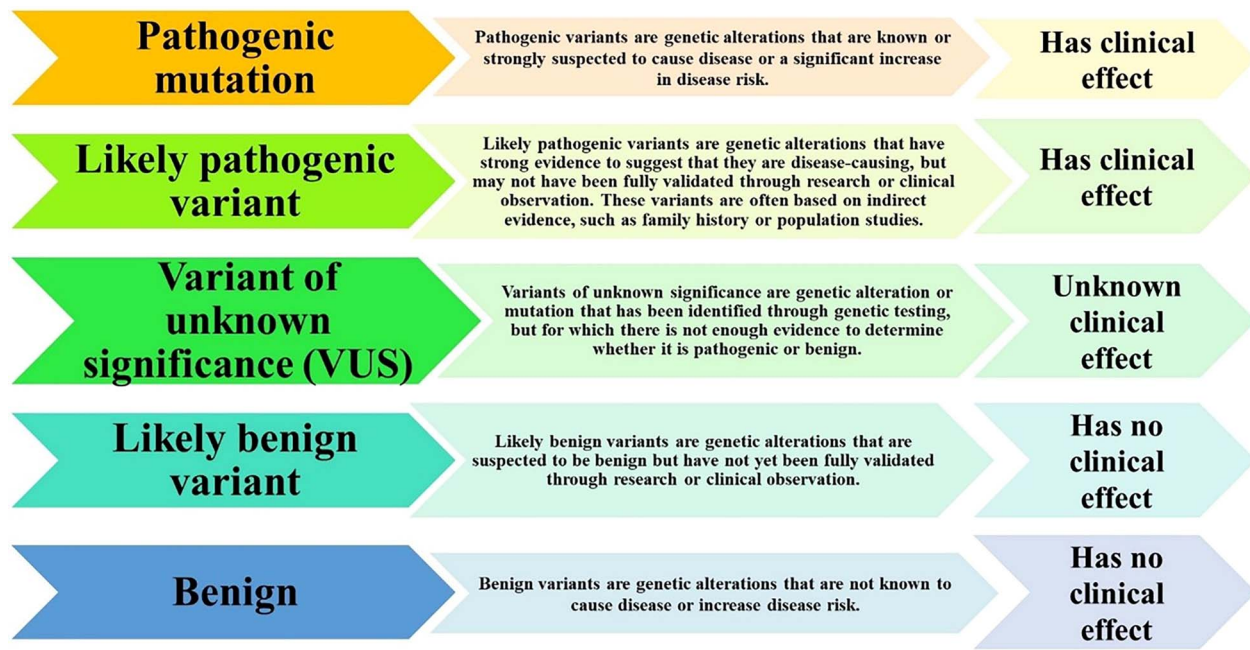
While previous efforts have integrated variant data from next-generation sequencing (NGS) for specific tools and workflow pipelines [23], focusing solely on sequence-related information has demonstrated limitations in accuracy [24]. Pathogenic variations can alter a protein's structural features, particularly disulfide bond sites [25], and impact protein stability [26]. Understanding the effects of variants on protein stability is crucial, necessitating an exploration of a protein's structure, function and dynamic relationships. Despite the success of 3D structure classifiers [27–29], sequence-based methods outpace structure-based modeling methods in assessing the effects of single amino acid variants (SAVs).

This review aims to highlight the genetic variant databases and associated ML tools used for prediction of BC pathogenicity. We first summarize well-known BC gene variants including their location and function and the associated abnormality of each genetic variant. Following this, a review of databases used in this type of research is explored including the targeted disease, the accessibility, their advantages and disadvantages and the associated website of each database. An example of applying the databases for predicting BC pathogenicity is provided and discussion of the advantages and disadvantages of each database provided. Moreover, we describe the ML tools, the advantages and disadvantages and the algorithms underpinning each tool along with the tool accessibility.

## GENETIC VARIANTS DATABASES

The American College for Medical Genetics (ACMG) and the Association for Molecular Pathology (AMP) have issued guidelines to classify the challenging missense variants or variants of unknown significance (VUS) as pathogenic or benign [30]. This is a consequence of the rarity of missense variants and hence the lack of data-driven clinical evidence, such as segregation and case control. As the problem of VUS has grown over time, most clinical genetic tests reported in ClinVar [20] are VUSs, even among highly studied cancer predisposition genes like Breast Cancer 1 (BRCA1), Tumor Protein 53 (TP53) and Phosphatase and Tensin Homolog (PTEN) [20].

A missense single-nucleotide variant (SNV) can lead to an SAV, which is an alteration in the protein sequence. Missense variants that encode a single change in the amino acid sequence of an affected protein represent around 45% of the known disease variants associated with cancer [31–33]. SNVs can be synonymous, non-synonymous or stop gain change. Each type alters the function of the protein differently. Indels and structural variations are also variants that result in altered protein function. Another



**Figure 2.** The classification of variants, their definition and their clinical effect.

division of the variants type is based on the type of cells, which can be either germline or somatic [34]. The types of variants are summarized in Figure 1.

The distinction of a pathogenic SAV from a benign SAV is critical for improving knowledge of the relationship between genes and diseases in the post-genomic age and facilitating the identification of innovative treatment methods for complex disorders. The accurate classification of a genetic variant effect on diseases is challenging to attain regardless of the abundance of the accumulated genetic variants data over the past few decades. Most existing functional impact prediction software for amino acid changes considers that protein sequences have survived natural selection among recognized living species. As a result, evolutionarily conserved amino acid locations across various species are considered functionally significant, while those found at conservation sites are considered to be harmful [35]. As per the ACMG guidelines, variants are classified into five categories based on their clinical effects. The classification of variants, their definition and their clinical effects are shown in Figure 2.

## BREAST CANCER

One of the most common tumors in the world and accounting for around 11% of all cancer mortality cases in women between 20 and 50 years is BC [5]. For women worldwide, it is responsible for more disability-adjusted life expectancy years than any other cancer. In any country, it can occur in women of any age group from puberty onward, and the risk increases with age. Therefore, there is an urgent need for a reliable and accurate system to aid in the early detection and diagnosis of BC.

### BC-related genes and variants

With the advances in technologies, specifically in the genomics area, many BC-associated genes have been identified in oncogenes and anti-oncogenes. Variants and aberrant amplifications are crucial to the development and growth of tumors. Family history and inherited genetic variants are one of the most critical risk factors associated with BC. Some variants in BC-related

genes are known to greatly affect the development of BC. On the other hand, many other genetic variants might affect BC but these are not yet clearly understood. Some gene variants are known to highly predispose women to the development of BC with some having penetrance reaching up to 80%. Most available pathogenicity prediction tools developed for BC focus mainly on well-known genetic variants such as *BRCA1*, *BRCA2*, *TP53* or *PTEN* variants. Training an ML tool on genes associated with BC can improve its prediction accuracy for BC cases. However, this needs to involve all known genes associated with BC development. Table 1 below summarizes currently well-known genes associated with BC that can be used as training data when developing a BC-specific pathogenicity prediction tool.

### Relevant genetic and genomic variants databases

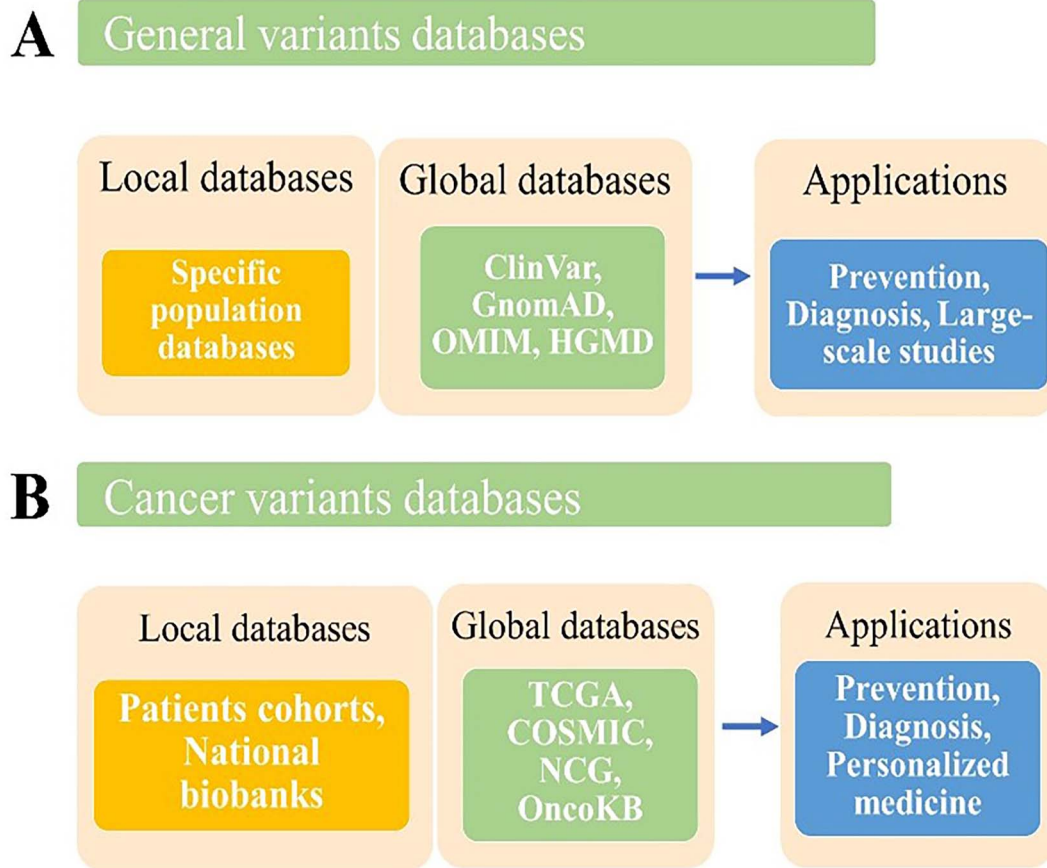
Recent advances in the field of molecular biology, coupled with the increased affordability of its associated techniques, have paved the way for the study of biological parameters, novel organisms and pathogens, as well as genetic diseases through the sequencing of genetic material. The vast amounts of data generated by these methods necessitate a high degree of expertise and computational power to process, identify and classify genetic variants that may provide scientifically valuable insights. Genomic studies have allowed us to uncover critical information and gain a better understanding of the molecular mechanisms underlying both our biology and various genetic diseases. By starting with the sequencing of small segments of genetic material and moving on to disease-specific gene panels and, more recently, whole exome and genome sequencing, we can, in some cases, trace the origins of a disease, enabling targeted therapy and significantly impacting the clinical decisions made for affected patients or their families [60, 61].

These studies allowed the creation of several databases and beyond, like The Cancer Genome Atlas (TCGA) [62], ClinVar [20] and The Catalogue of Somatic Mutations in Cancer (COSMIC) [19] and others, which provide us the curated data of the molecular alterations related to diseases and serve as a deposit for new studies. All these databases are major contributors to past and new

**Table 1:** Examples of genes associated with BC and their functions.

Gene	Full name and Location	Classification	Function	Abnormality after mutation	Reference
BRCA1	Breast Cancer 1 17q21	Tumor suppressor gene	Plays a role in DNA repair and maintenance of genomic stability.	Impairs the DNA repair function.	[36, 37]
BRCA2	Breast Cancer 2 13q12	Tumor suppressor gene	Plays a role in DNA repair, particularly in the repair of double-strand breaks.	Impairs the DNA repair function.	[38, 39]
HER2	Human Epidermal Growth Factor Receptor 2 17q12	Oncogene	Encodes a protein involved in cell growth and division.	Leads to uncontrolled cell growth and division.	[40]
EGFR	Epidermal Growth Factor Receptor 7p12	Oncogene	It is involved in cell growth and division.	Mutations cause constitutive activation of the EGFR receptor, leading to uncontrolled cell growth, division and progression of BC.	[41, 42]
c-Myc	Myc proto-oncogene protein8q24	Oncogene	Plays a critical role in the regulation of cell growth, differentiation and apoptosis.	Leads to dysregulated cell growth, impaired differentiation and decreased apoptosis, contributing to the development and progression of BC.	[43, 44]
Ras	Rat Sarcoma viral oncogene homolog- (Harvey) H-Ras - 11p15 (Kristen) K-Ras - 12p12 (Neuroblastoma) N-Ras - 1p22	Oncogene	Encode GTPases that play important roles in normal cell growth, differentiation and survival.	Leads to constitutive activation of the Ras protein, resulting in uncontrolled cell growth, impaired differentiation and resistance to apoptosis.	[45]
TP53	Tumor Protein 53 17p13.1	Tumor suppressor gene	Plays a critical role in maintaining genomic stability and preventing the development of cancer by promoting cell cycle arrest, DNA repair and apoptosis.	Leads to the accumulation of genetic damage and promoting the development and progression of BC.	[46, 47]
NME1	NME/NM23 nucleoside diphosphate kinase 1 17q21.3	Tumor suppressor gene	Plays a critical role in inhibiting tumor invasion and metastasis through its involvement in nucleotide metabolism, cell migration and signaling pathways.	Leads to the development and progression of BC by promoting tumor invasion and metastasis and is associated with a poorer prognosis.	[48, 49]
RB1	Retinoblastoma 1 13.2	Tumor suppressor gene	Regulates cell cycle progression, differentiation and apoptosis by controlling the activity of E2F transcription factors and other downstream targets.	Leads to uncontrolled cell proliferation, impaired differentiation and resistance to apoptosis.	[50, 51]
PTEN	Phosphatase and Tensin Homolog 10q23.3	Tumor suppressor gene	Regulates cell growth, proliferation and survival by negatively regulating the PI3K/Akt signaling pathway and promoting apoptosis and cell cycle arrest.	Leads to constitutive activation of the PI3K/Akt pathway, promoting uncontrolled cell growth, proliferation and survival.	[52, 53]
ATM	Ataxia Telangiectasia Mutated 11q22-q23	Tumor suppressor gene	Plays a critical role in detecting and repairing DNA damage, promoting cell cycle arrest and inducing apoptosis in response to genotoxic stress.	Impair the ability of cells to respond to DNA damage, leading to genomic instability and an increased risk of developing BC.	[54]
CDH1	Cadherin 1 16q22.1	Tumor suppressor gene	Encodes the E-cadherin protein, which plays a critical role in maintaining cell-cell adhesion, polarity and tissue architecture and regulating cell proliferation and differentiation.	Leads to reduced cell adhesion, impaired tissue integrity and enhanced cell motility and invasion.	[55]
FHIT	Fragile Histidine Triad 3p14.2	Tumor suppressor gene	Plays a critical role in regulating cell proliferation, DNA damage response and apoptosis, by promoting the cleavage of diadenosine triphosphate (Ap3A) and inhibiting signaling through the Wnt/ $\beta$ -catenin pathway.	Impair the ability of cells to respond to DNA damage and undergo apoptosis, promoting uncontrolled cell growth.	[56]
Maspin	Serpin Family B Member 5 18q21.33	Tumor suppressor gene	Regulates multiple cellular processes, including cell adhesion, migration, invasion, angiogenesis and apoptosis, by modulating signaling pathways involving integrins, growth factors and transcription factors.	Promote tumor growth, invasion and metastasis and is associated with a poorer prognosis in BC.	[57, 58]
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha 3q26.3	Oncogene	Encodes the p110 $\alpha$ subunit of phosphatidylinositol 3-kinase (PI3K), a critical signaling molecule that regulates cell growth, survival and metabolism, by activating the AKT/mTOR pathway and other downstream effectors.	Activate the PI3K pathway, leading to uncontrolled cell proliferation, survival and invasion.	[50]
CCND1	Cyclin D1 11q13	Oncogene	Encodes cyclin D1, a protein that promotes cell cycle progression by activating cyclin-dependent kinases and facilitating the transition from G1 to S phase and also has non-cycling functions in transcriptional regulation, cell migration and apoptosis.	Overexpression or amplification of CCND1 can drive excessive cell proliferation, survival and invasion.	[59]





**Figure 3.** Examples on cancer and general variants databases and applications.

studies and support variant classification [30, 63]. In 2015, several parameters were proposed by the ACMG [30] to be used to evaluate the pathogenicity of germline variants and one of the most widely applicable parameters is *in silico* analysis. This same analysis is also included in the guidelines for somatic variants as recommended in 2017 by the Association for Molecular Pathology, the American Society of Clinical Oncology and the College of American Pathologists [63] and more recently, in 2022, by the Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC) and Variant Interpretation for Cancer Consortium (VICC) [64].

Human variant databases usually have a specific scope and associated content. They can be used for predicting diseases [16] through supporting personalized medicine [17]. These databases have various limitations, including data structure compatibility and the variety of the data they hold in general. As a result, acquiring detailed information on a variation of interest is difficult [22]. Although use of several resources to analyze variant data has been explored [23], the data integration itself is largely for targeted tools and pipelines. Training and testing data are the most crucial elements for the success of any ML tool. The better the data used, the better the outcome. Different variant databases have different structures and datasets within them. Depending on the aim, the most appropriate database must be chosen. Figure 3 shows some examples of variants databases and their applications. The databases that are commonly used for the BC variant pathogenicity prediction tools are discussed below.

#### Human Gene Mutation Database (HGMD)

The Human Gene Mutation Database (HGMD) was initiated in 1996. It aimed to support the clinical study of variations in human

genes underlying genetic diseases [65, 66]. The HGMD aims to compile all known genetic variations that cause inherited disorders that have been reported in peer-reviewed journals including clinical genetic laboratories research. Over the last two decades, it has steadily gained a far more significant value as the principal unified repository for disease-related genetic germline variants. It has, for example, been used to enhance cancer prediction in high-risk hereditary BC families [67]. The HGMD provides a comprehensive set of published germline variants in genes that are thought to underlie or are closely associated with human-inherited disease. At the time of writing (December 2022), the HGMD comprised 234 987 publicly identified variants, with 117 744 privately identified variants from the HGMD Professional 2021.4. During the CAGI5 ENIGMA challenge, Color Genomics submitted four prediction sets with Learning from Evidence to Assess Pathogenicity (LEAP) [68, 69], an ML tool that predicts variant pathogenicity according to features including datasets from the HGMD and GnomAD databases. The overall performance accuracy achieved by LEAP was 83% [69].

#### ClinVar

ClinVar [70] is a free public human genetic variant collection comprising interpretations of their significance to diseases. It was released in 2013. The National Centre for Biotechnology Information (NCBI) maintains ClinVar within the National Library of Medicine (NLM) at the National Institutes of Health (NIH). Clinical testing laboratories, research laboratories, locus-specific databases, expert panels and other groups submit clinical significance information of variants or sets of variants to ClinVar [70]. ClinVar data were applied to a study by Metin and Pemra

[71] to assess the performance metrics of *in silico* pathogenicity methods on functional relevance of cancer variants obtained from ClinVar. They examined the pathogenicity predictions of cancer-related variant datasets of eight cancer types including BC retrieved from ClinVar using 13 different *in silico* tools. A combination of statistical performance metric analysis, prediction distribution frequency data and ROC curve analysis results have suggested that among all *in silico* prediction tools, the top three tools with the highest discriminatory power were found to be MutPred (AUC=0.677), MetaSVM (AUC=0.645) and Revel (AUC=0.637). ClinVar data were applied also in Lin et al. [72], where they identified BRCA1 VUSs from clinical sequencing data and wanted to interpret the clinical significance of such data. Several ML methods have been created to estimate the pathogenic hazards of variations of unknown significance. An optimized random forest algorithm outperformed the performance after benchmarking, and it was selected to predict BRCA1 VUSs from both the generated sequencing data and ClinVar data. A predicted pathogenicity of 6322 VUSs was obtained, of which 1593 variants were predicted to be pathogenic and 4729 were predicted to be benign [72].

### Catalogue Of Somatic Mutations In Cancer (COSMIC)

The Catalogue Of Somatic Mutations In Cancer (COSMIC) [19], launched in 2004, offers a collection of somatic variant data from various public sources through one standardized repository that makes it easy to be explored in various ways. COSMIC includes all forms of human cancers, from the most frequent to the extremely rare cancers, observed by clinicians possibly once or twice in a career. Data within COSMIC are collected from scientific publications of clinical, genetic and cancer-related research. COSMIC has developed into a large genome-wide system to investigate patterns of somatic variants in all cancer types. Moreover, recent studies have characterized specific variants in the evolution of genetic resistance to clinical therapeutics. The implementation of FATHMM-MKL (designed based on the characteristics of germline non-cancer variants) for predicting the pathogenic status of cancer somatic variants in the COSMIC dataset has shown good pathogenicity prediction results for BC [73, 74].

### The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) [62] and the International Cancer Genome Consortium (ICGC) were launched as the two major projects in 2005 and 2008, respectively. They were developed to use innovative genomic technologies including single-cell sequencing, whole genome and whole exome sequencing to improve the understanding of cancer genetics and create new methods of cancer treatment, diagnosis and prevention strategies. The National Institutes of Health initiated the TCGA Pilot Project to compile a comprehensive atlas of cancer genomic profiles. The TCGA is a public effort that intends to catalog and detect significant cancer-causing genomic changes in large cohorts of over 30 human malignancies utilizing modern genome sequencing techniques and integrated multi-dimensional analysis. These publicly available cancer genetic databases enable the advancement of diagnostic technologies, treatment guidelines and support [62, 75]. In a recent study, a total of 80 227 somatic SNVs from 976 patients were analyzed and the genomic features for 8647 somatic SNVs from 142 young patients (<45 years old at diagnosis) were identified. The data collected from the TCGA database included 6910 somatic SNVs from coding regions and 1737 somatic SNVs from non-coding regions of the genome [76].

### The Genome Aggregation Database (GnomAD)

The Genome Aggregation Database (gnomAD) [77] is one of the leading and most widely used collections of variants from synchronized sequencing data. To support quick and automatic variant analysis, the data are accessible through the online gnomAD browser. The Exome Aggregation Consortium (ExAC) dataset, the first significant compilation of existing sequence data from 60 000 individuals, was published in 2014 [78]. Mainly, gnomAD is generated using whole genome and whole exome sequencing data in addition to single-cell sequencing technologies. ExAC was renamed gnomAD after genome data were added, and it now contains variant data from more than 195 000 people. With more than 150 000 weekly page views, it is currently the most used reference population dataset. Using a non-Finnish non-cancer European population dataset as their control dataset, Rofes et al. [79] downloaded and filtered variants to identify predicted loss-of-function variants in *BRCA1-associated ring domain 1* (BARD1). Copy number variants screening was performed on the gnomAD SVs v2.1 dataset. This study showed results that support the role of BARD1 as a moderate-penetrance BC-predisposing gene and highlighted a strong association with triple-negative tumors [79].

### Network of Cancer Genes (NCG)

The Network of Cancer Genes (NCG) [80] is a comprehensive database released in 2010 that gathers a collection of curated cancer genes from cancer transcriptomic sequencing screens including next-generation sequencing, single-cell sequencing, whole exome and whole genome sequencing. The NCG is a freely available, manually curated repository of 2372 genes whose somatic modifications have known or predicted cancer driver roles. In 2018, the project reached its 6th release. The NCG genes were collected from 275 articles; 2 included known cancer genes and 273 included cancer sequencing screens from 34 905 cancer donors and various primary locations, covering more than 100 cancer types. In comparison to the previous version, this represents a content increase of more than 1.5-fold. Additionally, NCG annotates characteristics of cancer genes like duplicability, evolutionary origin, RNA and protein expression, interactions between miRNA and proteins and protein function and essentiality. The data from this database were not found in any pathogenicity prediction research, so it represents an exciting opportunity for the future [81].

### Online Mendelian Inheritance in Man (OMIM)

Online Mendelian Inheritance in Man (OMIM) [82] is a comprehensive and authoritative knowledge base of human genes and genetic disorders compiled to support human genetics research and education and support the practice of clinical genetics. It includes data from genome-wide association studies, next-generation sequencing, Sanger sequencing and others. OMIM is now distributed electronically by the NCBI. The Entrez suite of databases is combined with OMIM. Written and edited at Johns Hopkins University with input from scientists and doctors worldwide, OMIM is derived from biomedical literature. Each OMIM entry includes a full-text summary of a genetically determined phenotype and/or gene, as well as numerous links to other genetic databases, such as those for DNA and protein sequence, PubMed citations, general and locus-specific variant databases, HUGO nomenclature, MapViewer, GeneTests, patient support groups and a lot more. OMIM provides a gateway to the rapidly expanding body of knowledge in human genetics. OMIM

also has datasets on most cancer types, including BC that has not yet been used in any pathogenicity prediction tool training or testing.

### IntOGen-mutations

IntOGen-mutations [83] provides a resource for locating cancer drivers among various tumors that were identified using functional genomic analysis, whole exome and whole genome sequencing and so on. It can display the findings of the most recent large tumor somatic variant data sets that have undergone systematic analysis. It focuses on copy-number gains and losses and transcriptomic changes in tumors. The outcomes of tumor genome analyses conducted using various variant-calling workflows are integrated into the IntOGen-mutations database. To thousands of tumor genomes, it is scalable. Without the need to estimate the background variant rate, it offers a tool that identifies genes predisposed to accumulating variants with high functional effects. It also provides a tool that detects genes whose variants are highly functionally significant. Both tools look for signs of positive selection seen in genes whose variants are potential drivers of tumor formation. IntOGen-mutation data have not yet been used in research related to predicting the pathogenicity of BC-causing variants.

### cBio Cancer Genomics Portal (cBioPortal)

The open-source cBio Cancer Genomics Portal (cBioPortal) is a tool for viewing multi-dimensional cancer genomics data sets interactively [84]. It includes single-cell sequencing, whole exome and whole genome sequencing and other functional genomic assays data. Although open-source, germline datasets are not publicly accessible [85]. The cBioPortal has access to data from over 5000 tumor samples from 20 cancer studies [84]. The cBio Cancer Genomics Portal removes considerable barriers between complex genomic data and cancer researchers that want rapid and easy access to molecular profiles and clinical features from large-scale cancer genome studies. It helps researchers to get biological insights and clinical information by utilizing these large data sets. There are 15 initial TCGA data sets and 5 published data sets accessible on the cBioPortal. Based on the most recent TCGA production runs, provisional TCGA data sets are updated weekly, and the site is continuously updated when additional TCGA cancer types are introduced. Variant information is present in published data sets but not in tentative data sets. Variant data are made public and uploaded to the site once each cancer type within TCGA is completed and somatic variants are validated. The site also provides information on copy number changes, mRNA expression changes based on microarray and RNA sequencing, DNA methylation values, protein and phosphoprotein levels and variant data.

### DriverDBv2

DriverDBv2 [86] is an updated version of DriverDB. This is a database that includes over 6000 cases of whole exome and whole genome sequencing data, functional genomic assays and published bioinformatics techniques and annotation databases for driver gene/variant identification. The database provides two points of view, 'Cancer' and 'Gene', to help researchers visualize the connection between cancers and driver genes/variants. In the DriverDBv2 database, over 9500 cancer-related RNA-sequencing datasets and over 7000 exome-sequencing datasets were integrated from TCGA, ICGC and numerous published papers. Seven additional computational algorithms have been developed for driver gene identification and incorporated into the

analysis pipeline. Gu et al. applied FI-net and 22 other state-of-the-art tools to 31 datasets, including DriverDBv2 [87]. According to their comprehensive evaluation, FI-net outperformed other tools with results illustrating that FI-net could identify known and potential novel driver genes [87].

### OncoKB

OncoKB is an inclusive precision oncology knowledge database released in 2017 [88]. It provides comprehensive, evidence-based oncological somatic variants and structural changes knowledge found in patient tumors to support their therapy choices [88, 89]. It includes data generated through whole exome and whole genome sequencing, proteomics, immunohistochemistry and other functional genetic assays. OncoKB data are managed by a dedicated panel of clinicians and cancer biologists who evaluate and manage biomarker-associated investigational therapeutic strategies. OncoKB connects data on (Food and Drug Administration) FDA-approved treatments and investigational drugs undergoing clinical trial evaluation for biomarker-guided use. Additionally, it emphasizes unfavorable clinical findings to discourage the off-label use of costly targeted therapies that have been demonstrated to be ineffective in particular variational contexts. An interactive website and the cBioPortal for Cancer Genomics both offer access to OncoKB. By assisting doctors in finding potentially actionable variants to ensure that patients receive the proper remedies or are directed to the most pertinent clinical trials, a curated database like OncoKB can play a crucial role in helping to realize the promise of precision medicine [88].

### Functional Annotation of Somatic Mutations in Cancer (FASMIC)

Functional Annotation of Somatic Mutations in Cancer (FASMIC) [90] is a user-friendly, interactive and open-access web platform for comprehensive visualization and exploration of variant-associated data [90] collected from different genomic functional assays including next-generation sequencing, whole exome and whole genome sequencing. It includes modules such as brief description, 3D structures, literature, variant frequency, functional prediction and protein expression. To find a variant, users can first query its gene symbol and select the matched genes to show all related variants. All variations investigated are displayed in a tabular style, together with critical information for each variant, such as gene name, chromosomal location, amino acid change and functional annotation. A Function Prediction module gives function predictions generated by well-known computational techniques. Furthermore, a Protein Expression module provides extensive protein expression data of cell lines affected by variations compared with wild-type genes. This aids in understanding the unique functional effects of variations.

### Cancer Cell Line Encyclopedia (CCLE)

The Cancer Cell Line Encyclopedia (CCLE) [91] is a collection of 947 human cancer cell lines' genomic functional assays including whole exome and genome sequencing, gene expression, genomic copy number and massively parallel sequencing big-scale genomic datasets, as well as pharmacologic assays of 24 drugs across over 500 of these lines [91]. The CCLE encompasses 36 tumor types with several genomic technology platforms used for characterizing cell lines. The variational status of over 1600 genes was determined by targeted massively parallel sequencing, followed by removal of variants likely to be germline events. 392 recurrent variants affecting 33 known cancer genes were assessed

by mass spectrometric genotyping. DNA copy number was measured using a high-density single-nucleotide polymorphism array. Eventually, mRNA expression levels for each of the lines were determined. These results were also utilized to validate cell lines. In a drug response prediction study and through leave-one-out cross-validation and cross-classification on independent datasets, it was shown that using this dataset for prediction leads to an accurate and reproducible classification of sensitive and resistant cell line–drug pairs with a high degree of accuracy [92].

## Comparison

Several available databases have been developed for cancer-causing gene identification. The differences in the data structures and nature of the data types in each database along with diversity of curation information give different results when comparing these resources using ML tools. Moreover, some databases like the CCLE, COSMIC and others demonstrate functional information regarding the variant and its effect on the interaction of the drug with its ligand; based on that, personalized treatment for each patient's variant can be established. The personalized treatment can be either a new drug or natural product that is found to bind perfectly to the mutated ligand or repurposed drug, which is any FDA-approved drug that was not initially indicated to treat the disease but is found to be perfectly act on the mutated ligand. Our goal is to choose the most suitable data sources for a given tool to predict the pathogenicity of variants. Some databases were not previously used to train or test any BC pathogenicity prediction tools; however, they are good candidates for future BC-specific tools training and testing. Table 2 summarizes the databases of variants, the variant type and the accessibility and the location (website) of each database. Additionally, an application of different cancer-related databases in BC pathogenicity prediction is provided in Table 3 with the advantages and disadvantages of each database summarized in Table 4.

## ML TOOLS FOR PREDICTION OF BC PATHOGENICITY

The 'gold standard' prediction of BC pathogenicity as per the ACMG guidelines involves screening procedures consisting of clinical evaluation, radiological imaging and pathological testing [93]. Due to the fact that the traditional gold standard of classification is expensive, human invasive and intensive, some highly accurate prediction tools like SIFT can be used to help in pathogenicity classification. Additionally, new ML tools can be used to serve a similar purpose based on model creation and extensive training and validation. In the training and testing stage, a given ML model makes predictions using input data comprising known/confirmed BC pathogenicity data and benign data [94]. Pre-processing, feature selection and extraction and classification are key elements of ML [95]. The feature extraction part of an ML tool is crucial for cancer diagnosis and prediction. The workflow of the pathogenicity prediction research using ML is shown in Figures 4 and 5.

Many ML tools have been developed and applied to predict the potential pathogenic effect of variants. Some of these tools were developed explicitly for given diseases, while others have been developed to be general purpose. In this work, we consider 14 ML-based and 2 non-ML-based pathogenicity prediction tools, which we discuss below. The non-ML tools were added to be able to compare between the ML- and non-ML-based tools performance. We provide a description of each tool, the type of application, the advantages and disadvantages, the algorithm used in the

development of the tool and the reliability and tool links for each of the tools.

## Combined Annotation-Dependent Depletion (CADD)

Combined Annotation-Dependent Depletion (CADD) [97] is a free, commonly used pathogenicity prediction tool that uses a logistic-regression ML model to categorize causal variants in genetic analysis, with a specific focus on highly penetrant contributors to severe Mendelian disorders. It was originally trained on various datasets from different databases including gnomAD, ClinVar and others. It offers an integrative annotation built from more than 60 genomic features and can score SNVs and short insertions and deletions anywhere in the reference assembly. The ML model CADD uses is trained on a binary distinction between simulated *de novo* variants and fixed variants in humans. The utility of the CADD score was recently reported to rank pathogenicity as C-scores ranging from 1 to 99 for deleterious variants. Using C-scores, Nakagomi et al. attempted to constitute a classification system for *BRCA1* and *BRCA2* variants of uncertain significance. It was found that CADD can classify *BRCA1* and *BRCA2* variants and select patients for further segregation studies [97].

## Polymorphism Phenotyping v2 (PolyPhen-2)

Polymorphism Phenotyping v2 (PolyPhen-2) [98] is an ML tool that is used to predict the possible impact of amino acid substitutions on the structure and function of a human protein. It was trained on a variation of databases including UniProt, NCBI RefSeq, sequence alignment and others. It uses a combination of physical and comparative considerations to make predictions. PolyPhen-2 uses eight sequence-based and three structure-based predictive features to predict the effect of a mutation on protein function. These features are selected automatically by an iterative greedy algorithm, which iteratively selects the features that improve the prediction accuracy the most. The algorithm is designed to consider both the overall accuracy of the predictions and the balance between sensitivity and specificity. The distance between the protein containing the first variation from the human wild-type allele and the human protein and whether the mutant allele originated at a hypermutable site are the characteristics that characterize how well two human alleles fit into the pattern of amino acid replacements within the context of multiple sequence alignment of homologous proteins. Using a clustering algorithm, the alignment pipeline chooses the set of homologous sequences to be examined before building and fine-tuning their alignments [98]. The functional significance of an allele replacement is predicted from its individual features based on a Naïve Bayes classifier. In terms of accuracy, [99, 100] reported the performance of PolyPhen-2 for predicting the functional effects varied across a clinical dataset of *BRCA1* and *BRCA2* missense variants. The absence of consistency in prediction outcomes limit the clinical application in classifying pathogenic VUSs identified through molecular testing of *BRCA1* and *BRCA2* [101].

## Fathmm-MKL

Fathmm-MKL [74] is an ML tool that is used to predict the functional effects of missense variants in a protein by combining sequence conservation within hidden Markov models (HMMs), indicating the alignment of homologous sequences and conserved protein domains. Pathogenicity weights are used for the overall tolerance of the protein to variants. Fathmm-MKL is trained on an integration of databases including



**Table 2:** The summary of the databases of cancer-related variants, the variant type, the accessibility and the location of each database.

Database	Full form name	General description	Targeted disease	Variants type	Website
HGMD	Human Gene Mutation Database	Is a repository of inherited variant data useful for medical research, genetic diagnosis and next-generation sequencing studies.	General	Somatic and germline	<a href="https://www.hgmd.cf.ac.uk/ac/index.php">https://www.hgmd.cf.ac.uk/ac/index.php</a>
ClinVar	Clinical significance variants	Is a freely available archive for interpretation of the clinical significance of variants for reported conditions.	General	Somatic and germline	<a href="https://www.ncbi.nlm.nih.gov/clinvar/">https://www.ncbi.nlm.nih.gov/clinvar/</a>
COSMIC	Catalogue of Somatic Mutations in Cancer	Is a database of information about somatic variants in cancer obtained from curating relevant literature and high-throughput sequencing data generated by the Cancer Genome Project and others.	Cancers	Somatic	<a href="https://cancer.sanger.ac.uk/cosmic">https://cancer.sanger.ac.uk/cosmic</a>
TCGA	The Cancer Genome Atlas	Is a project to identify the complete set of DNA changes in many different types of cancer.	Cancers	Somatic and germline	<a href="https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga">https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga</a>
GnomAD	The Genome Aggregation Database	Is a resource developed by an international coalition of investigators to aggregate and harmonize exome and genome sequencing data from a wide variety of sequencing projects with summary data for the broader scientific community.	General	Germline	<a href="https://gnomad.broadinstitute.org/">https://gnomad.broadinstitute.org/</a>
NCG	Network of Cancer Genes	Is a manually curated repository of genes whose somatic modifications have known or predicted cancer driver roles.	Cancers	Somatic	<a href="http://ncg.kcl.ac.uk/">http://ncg.kcl.ac.uk/</a>
OMIM	Online Mendelian Inheritance in Man	Is a continuously updated catalogue of human genes and genetic disorders and traits, with particular focus on the gene–phenotype relationship.	General	Allelic	<a href="https://www.omim.org/">https://www.omim.org/</a>
IntOGen	Integrative Onco Genomics	Is a framework for automatic and comprehensive knowledge extraction based on variant data from sequenced tumor samples. The framework identifies cancer genes and pinpoints their putative mechanism of action across tumor types.	Cancers	Somatic	<a href="https://www.intogen.org/about">https://www.intogen.org/about</a>
CBioPortal	The cBio Cancer Genomics Portal	Is an exploratory analysis tool for exploring large-scale cancer genomic data sets from large consortium efforts, like TCGA, as well as publications from individual labs.	Cancers	Somatic and germline	<a href="https://www.cbioportal.org/">https://www.cbioportal.org/</a>
DriverDB	Driver Database	Is a cancer omics database that integrates somatic variants, RNA expression, miRNA expression, methylation, copy number variation and clinical data with annotation and published bioinformatics algorithms.	Cancers	Somatic	<a href="http://driverdb.tms.cmu.edu.tw/">http://driverdb.tms.cmu.edu.tw/</a>
OncoKB	Oncology Knowledge Base	Is a comprehensive and curated precision oncology knowledge base that offers oncologists detailed, evidence-based information about individual somatic variants and structural alterations in patient tumors with the goal of supporting optimal treatment decisions.	Cancers	Somatic	<a href="http://oncokb.org">http://oncokb.org</a>
FASMIC	Functional Annotation of Somatic Mutations in Cancer	Is a comprehensive database for understanding the functional impact of somatic variants in cancer. It provides functional annotations with protein expression, variant frequency, 3D structures, function prediction and literature to help researchers explore variant details.	Cancers	Somatic	<a href="https://bioinformatics.mdanderson.org/public-software/fasmic/">https://bioinformatics.mdanderson.org/public-software/fasmic/</a>
CCLE	Cancer Cell Line Encyclopedia	Is a database of gene expression, genotype and drug sensitivity data for human cancer cell lines.	Cancers	Somatic and germline	<a href="https://sites.broadinstitute.org/ccle/datasets">https://sites.broadinstitute.org/ccle/datasets</a>

functional annotations from ENCODE with nucleotide-based sequence conservation measures when assessing the functional consequences of coding and non-coding variants in addition to others. It was observed that Fathmm-MKL had improved performance when compared with other algorithms like CADD

when predicting the functional impact of SNVs [74]. Nono *et al.* have shown that Fathmm-MKL effectively predicted the pathogenicity of BC-causing gene variants with a Pearson's correlation coefficient of 0.80, outperforming other tools used in that research [73].

**Table 3:** The application of different cancer-related databases in BC pathogenicity prediction.

Database	Application example	Reference
HGMD	Color Genomics by Lai <i>et al.</i> submitted four sets of predictions using LEAP, a machine learning framework that predicts variant pathogenicity according to features based on training datasets from the HGMD.	[68]
ClinVar	Lin <i>et al.</i> identified BRCA1 VUSs from clinical sequencing data. 1593 VUSs were predicted to be pathogenic, and 4729 VUSs were predicted to be benign. Yazar <i>et al.</i> used a combination of statistical performance metric analysis, prediction distribution frequency data and ROC curve analysis results have suggested that among all <i>in silico</i> prediction tools, the top three tools with the highest discriminatory power were found to be MutPred (AUC=0.677), MetaSVM (AUC=0.645) and Revel (AUC=0.637).	[71, 72]
COSMIC	FATHMM-MKL was used for predicting the pathogenic status of cancer somatic variants in the COSMIC dataset. It was shown to have good prediction results for BC pathogenicity.	[73]
TCGA	Feizi <i>et al.</i> applied various models to predict the pathogenic status of somatic variants identified in young BC patients from TCGA-BRCA studies. The results indicated that using their model predicted 1853 positive SNVs (out of 6910) from the TCGA-BRCA dataset.	[76]
GnomAD	Rofes <i>et al.</i> used the gnomAD non-Finnish European population, non-cancer dataset as a control population for their study. This study showed results that support the role of BARD1 as a moderate-penetrance BC-predisposing gene and highlight a strong association with triple-negative tumors.	[79]
DriverDB	Gu <i>et al.</i> applied FI-net and other 22 state-of-the-art tools to 31 datasets including DriverDBv2. According to the comprehensive evaluation, FI-net outperformed the other tools. Furthermore, the results illustrated that FI-net could identify known and potential novel driver genes.	[87]
CCLL	In a drug-response prediction study and through leave-one-out cross-validation and cross-classification on independent datasets, it was shown that using this dataset in the prediction leads to accurate and reproducible classification of sensitive and resistant cell line-drug pairs with a high degree of accuracy.	[92]

### Rare Exome Variant Ensemble Learner (REVEL)

Rare Exome Variant Ensemble Learner (REVEL) is an ML ensemble tool used for predicting the pathogenicity of missense variants based on several other tools including MutPred, FATHMM, VEST, PolyPhen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP and phastCons. REVEL was trained with recently discovered pathogenic and rare neutral missense variants and excluded those used previously in training the original (individual) tools; this makes up the huge volume of the data used to train REVEL overall. REVEL performed very well in predicting the pathogenicity of variants compared with individual tools [102]. Although REVEL was not initially developed for predicting BC pathogenic variants, it has shown good performance with an area under the curve (AUC) of 0.79, which is one of the highest accuracy values compared with tools not designed specifically for BC [103].

### CScape

CScape [104] is an ML-based tool for predicting the probability of a variant to drive cancer. It was trained using datasets from COSMIC and 1000 Genomes Project databases. CScape outperforms alternative tools on somatic variants, reaching 91% accuracy in coding regions and 70% in non-coding regions. Using thresholds to separate high-confidence predictions can increase accuracy. A statistical method was used to distinguish the coding from the non-coding regions of the cancer genome, which tends to cluster in genomic regions where optimistic predictions are made to distinguish between recurrent and rare variants in the human cancer genome in advance [104]. CScape-somatic [105] is an integrative classifier tool that is used to predictively discriminate between recurrent and rare variants in the human cancer genome. It was trained on datasets from the COSMIC database and the International Cancer Genome Consortium Data Portal. This tool is designed to work with somatic point variants in both coding and non-coding regions of the genome. It uses only cancer genome data to examine the difference between rarely occurring and frequently occurring somatic single-point variants in the

human cancer genome. The authors of this tool have shown that this type of predictive differentiation can offer a fresh perspective and potentially a more precise prediction in both the coding and non-coding regions of the cancer genome. It's important to note that this tool is focused on somatic mutations, which are mutations that occur in cells that are not germ cells and that are not passed down to the next generation. This is different than germline mutations that are present in every cell of the body and are inherited from a parent. When tested on somatic variants, CScape-somatic outperforms rival tools, achieving balanced accuracy in coding areas of 74% and non-coding regions of 69%. Using thresholds to extract high-confidence predictions can increase accuracy [105].

### DeepDriver

DeepDriver [106] is an ML-based tool based on deep neural networks that performs convolution of variant-based features of genes and their neighbors in similarity networks. It was suggested that similarity networks and attributes that describe the functional impact of variants might be used to determine driver genes. A convolutional neural network trained using a variant-based feature matrix built based on the topological structure of a similarity network specifically predicts putative driver genes. This tool is trained on different datatypes including gene expression data from the NCI Genomic Data Commons and functional annotations from COSMIC. The technique takes advantage of the similarities between gene expression patterns and the functional effects of variants simultaneously. This makes it possible to combine two types of data and increase prediction accuracy. The technology improves the prediction of driver genes by enabling the convolutional neural network to learn information from variant data and similarity networks simultaneously [106].

### DNA-repair Associated Breast Cancer (DrABC)

DNA-repair Associated Breast Cancer (DrABC) [107] is another DL-based tool that enhances the accuracy of identifying germline pathogenic variants (GPVs) carriers in cancer predisposition genes

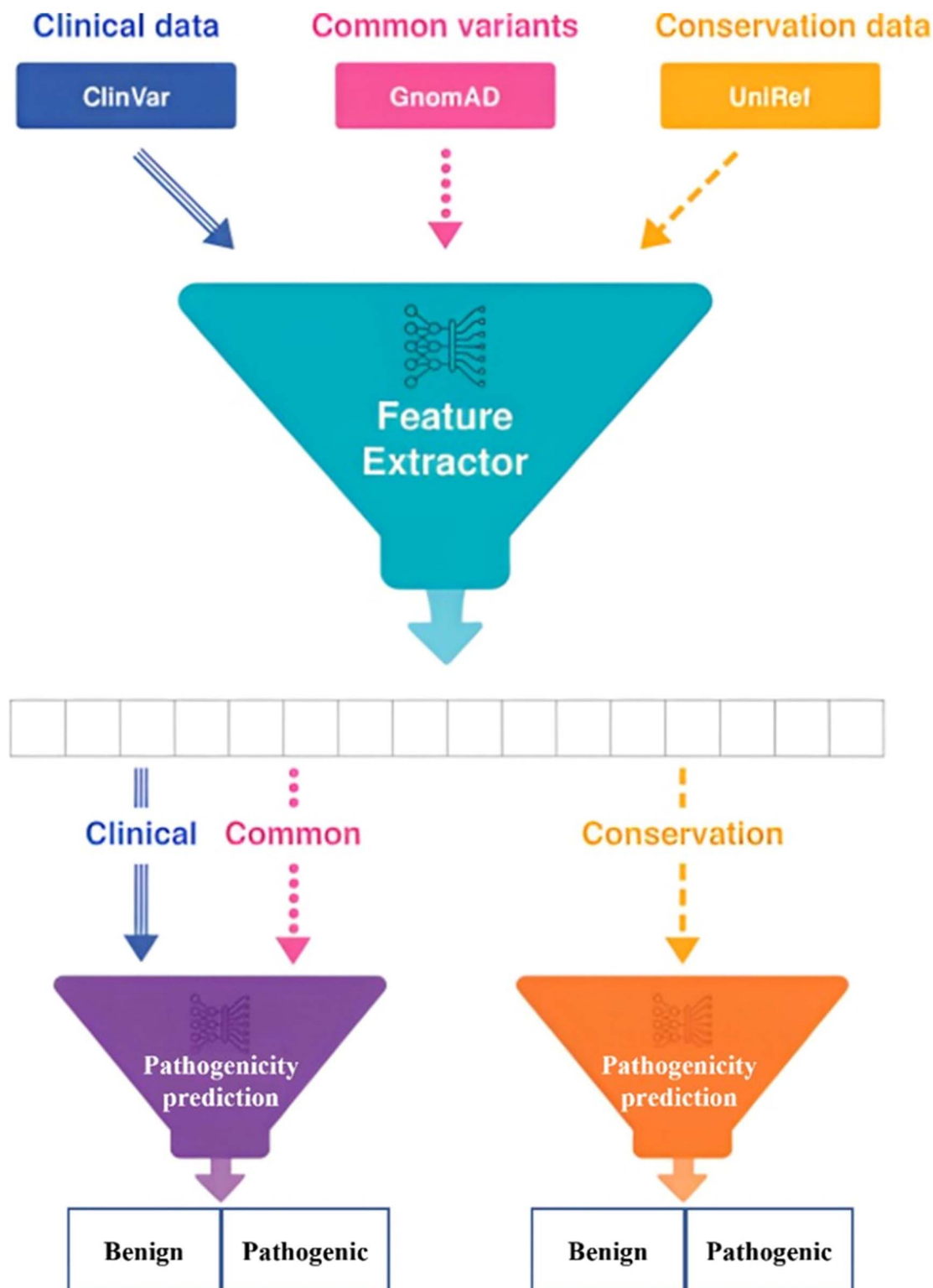
**Table 4:** The advantages and disadvantages of the presented databases.

Database	Advantages	Disadvantages
HGMD	<ul style="list-style-type: none"> <li>Comprehensive for all disease-causing variants.</li> <li>Provides variant-specific links to several other databases.</li> </ul>	<ul style="list-style-type: none"> <li>Includes only a single reference for each variant.</li> <li>Includes only disease-causing variants for general diseases.</li> </ul>
ClinVar	<ul style="list-style-type: none"> <li>Comprehensive for all known disease-causing and non-disease-causing variants.</li> </ul>	<ul style="list-style-type: none"> <li>Includes variants regardless of association with disease.</li> <li>Provide access to all observed variants but may not be supported by peer-reviewed literature.</li> </ul>
COSMIC	<ul style="list-style-type: none"> <li>Accurate and consistent data.</li> <li>Actionability functionality allows users to search drugs that target somatic variants at all stages of drug development, including those still in development, in clinical trials or that have been repurposed.</li> </ul>	<ul style="list-style-type: none"> <li>Manually curated, which is time-consuming and not rapidly modified.</li> </ul>
TCGA	<ul style="list-style-type: none"> <li>Provides a large number of cancer-specific samples.</li> <li>Offers multiple data platforms for the same sample.</li> <li>Offers unified data generation and low-level analysis.</li> </ul>	<ul style="list-style-type: none"> <li>The clinical data are spotty as almost all the samples are primarily untreated, without any response data and short follow-up.</li> <li>There are no immune-oncology data.</li> <li>Samples in the TCGA project are all fresh frozen samples, which are not commonly used in clinical settings.</li> </ul>
GnomAD	<ul style="list-style-type: none"> <li>GnomAD's predecessor, the Exome Aggregation Consortium (ExAC) database, lies in capturing sequencing data representing diverse European and non-European ancestries at a larger scale compared with previous sequencing studies.</li> </ul>	<ul style="list-style-type: none"> <li>Many populations are underrepresented.</li> <li>Some variants are somatic clonal variants.</li> <li>Not everyone in gnomAD is healthy and young.</li> </ul>
NCG	<ul style="list-style-type: none"> <li>It has cancer-specific variants.</li> <li>It incorporates information about genes with a known or anticipated significance as cancer drivers (predisposition).</li> </ul>	<ul style="list-style-type: none"> <li>It requires the use of <i>ad hoc</i> tools for data organizing and mining.</li> </ul>
OMIM	<ul style="list-style-type: none"> <li>Continuously updated.</li> <li>It unravels the complex relationships between genes and disease.</li> </ul>	<ul style="list-style-type: none"> <li>Only few non-protein-coding genes variants are included.</li> </ul>
IntOGen	<ul style="list-style-type: none"> <li>It has cancer-specific variants.</li> </ul>	<ul style="list-style-type: none"> <li>Only contain somatic variants.</li> </ul>
CBioPortal	<ul style="list-style-type: none"> <li>It integrates multiple cancer genomics projects.</li> <li>It enables the users to analyze complex data sets and translate into biologic insights and immediate clinical applications.</li> </ul>	<ul style="list-style-type: none"> <li>It has potential bias to estimate the relative proportion of germline variants, <i>de novo</i> variants and rare mutated alleles in a sample.</li> </ul>
DriverDB	<ul style="list-style-type: none"> <li>It incorporates large-scale data mining using many algorithms and then presents summarized driver genes with different kinds of aspects for variant visualization.</li> </ul>	<ul style="list-style-type: none"> <li>It uses tools like SIFT and PolyPhen to calculate scores, although they are not cancer-specific tools, so the results might not be reliable.</li> </ul>
OncoKB	<ul style="list-style-type: none"> <li>It is oncologist-oriented with evidence-based information about individual somatic variants and structural alterations present in patient tumors to support optimal treatment decisions.</li> </ul>	<ul style="list-style-type: none"> <li>As it oncologist specialized, other users might not understand the data.</li> </ul>
FASMIC	<ul style="list-style-type: none"> <li>It provides a comprehensive database for functional impact of somatic variants in cancer.</li> </ul>	<ul style="list-style-type: none"> <li>It does not cover germline variants.</li> </ul>
CCLE	<ul style="list-style-type: none"> <li>It includes data on gene variants, RNA splicing, DNA methylation, histone H3 modification and microRNA expression.</li> </ul>	<ul style="list-style-type: none"> <li>The effects of variant in the cell line and in humans might be different.</li> </ul>

(CPGs). It can locate GPVs and CPGs among BC patient-centered different endophenotypes with GPVs in genes engaged in homologous recombination and other DNA repair pathways. It was trained on a Chinese-specific discovery cohort. Lui *et al.* evaluated a multi-center cohort of 3041 female Chinese BC patients who underwent multi-gene genetic testing. Incorporating the detailed phenotypes of numerous cancer types and their family histories. A phenotype-driven prediction model based on a hierarchical neural network architecture was developed to recognize hereditary BC by considering the distinct endophenotypes linked to various CPGs in BC patients. When used to identify GPV carriers among Chinese BC patients, the model performed better than expected [107]. However, such tools are specific to a single disease instead of dealing with all diseases.

## RENOVO

RENOVO [108] is a computational ML-based tool that uses a random forest algorithm to classify genetic variants as pathogenic or benign based on publicly available information. It is trained on a set of pathogenic and benign variants from the ClinVar database. It has been validated on additional datasets, including unreported variants validated either through expert agreement (ENIGMA) or laboratory-based functional assays of BRCA1/2. The tool uses feature classifications based on the same guideline recommendations as other existing tools, but it outperforms these other tools on all datasets. This is important as it provides a validated tool to reduce the fraction of uninterpreted or misinterpreted variants, an unmet need in modern clinical genetics. RENOVO can achieve high performance by using a random forest algorithm.



**Figure 4.** Applying ML in the pathogenicity prediction research. This figure was modified from Won et al. [96].

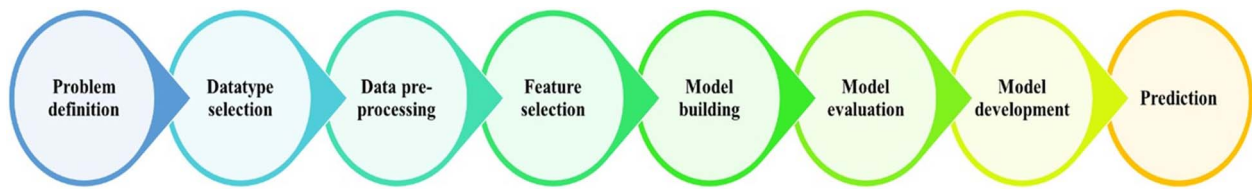
This ML algorithm can learn from large amounts of data and identify complex relationships between input features and output labels. It can help improve the interpretation of genetic variants in the clinical setting, which can help diagnose and manage genetic diseases [108].

### Supervised machine learning framework (SVFX)

Supervised machine learning framework SVFX [109] is an ML-based tool to score the pathogenicity of somatic and germline

structural variants (SVs). SVFX was trained on datasets from the Pan-Cancer Analysis of Whole Genomes (PCAWG) Project, Genome Sequencing Program (GSP), ClinVar database, gnomAD and 1000 Genomes Project. SVs play a critical role in many diseases, but limited approaches are available for interpreting and prioritizing these variants [110]. SVs cause more substantial variation in an individual genome at the nucleotide level than other variants. Still, they should be more noticed due to the technical challenges associated with their detection and analysis [110]. To





**Figure 5.** The main workflow of the pathogenicity prediction research using ML.

address this challenge, the authors of SVFX developed a new framework that utilizes tissue-specific genomic and epigenomic features to score the pathogenicity of SVs [109]. The framework was trained using SV call sets in diseased and healthy individuals and included genomic, epigenomic and conservation-based features. SVFX was applied to SVs in cancer and other diseases and achieved high accuracy in classifying pathogenic SVs. The predicted pathogenic SVs in cancer cohorts were found to be enriched among known cancer genes and many cancer-related pathways. SVFX is a valuable tool for identifying and interpreting structural variants, which can provide a more comprehensive understanding of the molecular mechanisms of various diseases. It can help identify potential driver mutations in cancer and other diseases, aiding in developing new treatments [109].

### Aljarf et al.

Aljarf et al. [103] developed an ML-based tool for evaluating the functional impact of single-point missense variants in the BRCA1 and BRCA2 genes. The tool uses supervised ML, which is a reliable approach for categorizing missense variants in a gene with given clinical effects. It was trained on evolutionary conservation, missense variant prediction models from dbNSFP, physicochemical properties and changes in post-translational modifications. The tool is designed to be both gene-specific for BRCA1 and BRCA2 and also a generic tool for evaluating missense variants in other genes. The authors anticipate that this *in silico* saturation mutagenesis tool will be valid and reliable for detecting variants of uncertain significance (VUS) and providing precise functional estimations for newly discovered variants [103]. Additionally, the enhanced prediction performance of the tool could assist researchers in classifying possible single-nucleotide variants (SNVs) in BRCA1 and BRCA2 for further exploration and validation. The tools were validated using 10-fold cross-validation, and the final tool models achieved a Matthew's Correlation Coefficient of up to 0.98. It is assumed that this predictive tool can be an effective tool for guiding the analysis of newly discovered variants and prioritizing variants for experimental validation. It can provide insights into understanding and interpreting the functional outcomes of missense variants in these genes. This tool can be a valuable resource for researchers and clinicians as it can assist in the identification of potential disease-causing variants in BRCA1 and BRCA2 genes, which are associated with an increased risk of breast and ovarian cancer [103].

### MutPred and MutPred2

MutPred is a random forest-based ML tool that depends on sequence, conservation, structural and functional characteristics to predict a variant's pathogenicity classification [111]. MutPred2 is a neural network ensemble tool with an expanded feature set that has been trained on a much larger and more heterogeneous dataset acquired from HGMD, SwissVar, dbSNP and others [112]. MutPred2 was run based on two approaches: with and without considering gene families in training. These characteristics

itemize proteins in the human and mouse genomes at several levels of sequence identity to the protein in which the variant is detected. These features were informally referred to as 'homolog counts'. The only inputs needed for MutPred and MutPred2 are a protein sequence and an amino acid substitution as input and output scores between zero (benign) and one (pathogenic). Both tools provided accurate predictions for BRCA1 however MutPred outperformed MutPred2 for BRCA2. Both tools performed similarly when the 'probably benign' were excluded. This was possibly as a result of selection of the MutPred2 model that included protein-level homolog counts as features [112].

### Learning from Evidence to Assess Pathogenicity (LEAP)

Learning from Evidence to Assess Pathogenicity (LEAP) [68] is an ML-based pathogenicity prediction tool. It was trained on missense variants detected and classified during routine clinical testing at Color Genomics. Manual variant classification uses various underlying data types, such as functional prediction, splice prediction, evolutionary conservation, population frequency, protein domain, co-occurring pathogenic (P/LP) variants and individual and family medical histories. LEAP prioritizes the evidence and weights it according to how it contributes to predictions based on the work of different scientists. LEAP's prediction performance was assessed with growing evidence based on several model types and evaluations of numerous genes and disease areas. Its value as a tool for clinical interpretation was explored based on the Critical Assessment of Genome Interpretation (CAGI5) ENIGMA Consortium [69] who held a blind prediction challenge. Variations of LEAP placed first, second, third and fourth against competing models that were either published or newly developed. This was the first external validation of LEAP's performance. Apart from excluding any inputs that are not readily available to the general public, LEAP2 acts as a control and is equal to LEAP1 in terms of pathogenicity estimation including use of data from HGMD. Random forest is used in LEAP3 as opposed to regularized logistic regression. Instead of a two-class model (Benign, Pathogenic), LEAP4 employs a three-class regularized logistic regression model (Benign, VUS, Pathogenic).

### LYRUS

LYRUS [113] is an ML tool based on the XGBoost classifier developed to predict the pathogenicity of SAVs. It was trained based on variants collected from the ClinVar database. Most variants in the human genome come from SAVs. Understanding the genomic architecture of complex diseases can be obtained by identifying pathogenic SAVs. Most methods for predicting the pathogenicity or functional impacts of SAVs rely on either structural or sequencing data. LYRUS combines five sequence-based, six structure-based and four dynamics-based features. Uniquely, LYRUS integrates the variation number, a recently suggested characteristic of sequence co-evolution. The ClinVar database's dataset of 4363 protein structures corresponding to 22 639 SAVs were used to

**Table 5:** The summary of the application type, the programming language and the core algorithm realized by the tool.

Tool	Application type	Programming language	Algorithm	Reference
CADD	Web-based	Python	Logistic regression	[97]
PolyPhen-2	Web-based	Python	Naïve Bayes classifier	[98]
Fathmm-MKL	Web-based	Python	Multiple kernel learning	[74]
REVEL	Web-based	–	Ensemble method	[102]
CScape	Web-based	–	Integrative classifier	[104, 105]
DeepDriver	Downloadable code	Python	Convolutional neural network	[106]
DrABC	Web-based	Python	Deep learning	[107]
RENOVO	Web-based	Python + R	Random forest	[108]
SVFX	Downloadable code	Python	Supervised ML	[109]
Aljarf et al.	Private	–	Supervised ML	[103]
MutPred and MutPred2	Web-based	MATLAB	Random forest and neural networks	[112]
LEAP	Private	Python	Logistic regression and random forest	[68]
Lyrus	Downloadable code	Python	XGBoost	[113]

**Table 6:** Significance of each tool.

Tool	Number of citations (from Google scholar as of 5 September 2023)	Link of tool/source code	Reference
CADD	17	<a href="http://cadd.gs.washington.edu/">http://cadd.gs.washington.edu/</a> . <a href="https://github.com/kircherlab/CADD-scripts">https://github.com/kircherlab/CADD-scripts</a>	[97]
PolyPhen-2	3297	<a href="http://genetics.bwh.harvard.edu/pph2/">http://genetics.bwh.harvard.edu/pph2/</a>	[98]
Fathmm-MKL	602	<a href="http://fathmm.biocompute.org.uk/">http://fathmm.biocompute.org.uk/</a> <a href="https://github.com/HAShihab/fathmm-MKL">https://github.com/HAShihab/fathmm-MKL</a>	[74]
REVEL	1429	<a href="https://sites.google.com/site/revelgenomics/about">https://sites.google.com/site/revelgenomics/about</a>	[102]
CScape	48	<a href="http://cscape.biocompute.org.uk/">http://cscape.biocompute.org.uk/</a>	[104, 105]
DeepDriver	16	<a href="http://cscape-somatic.biocompute.org.uk/">http://cscape-somatic.biocompute.org.uk/</a>	
DrABC	67	<a href="https://github.com/luoping1004/deepDriver">https://github.com/luoping1004/deepDriver</a>	[106]
	2	<a href="http://gifts.bio-data.cn/#/">http://gifts.bio-data.cn/#/</a>	[107]
RENOVO	13	<a href="https://github.com/zhq921/DrABC">https://github.com/zhq921/DrABC</a> <a href="https://bioserver.ieo.it/shiny/app/renovo">https://bioserver.ieo.it/shiny/app/renovo</a> <a href="https://github.com/mazzalab-ieo/renovo">https://github.com/mazzalab-ieo/renovo</a>	[108]
SVFX	22	<a href="https://github.com/gersteinlab/SVFX">https://github.com/gersteinlab/SVFX</a>	[109]
Aljarf et al.	3	Private	[103]
MutPred and MutPred2	45	<a href="http://mutpred.mutdb.org/">http://mutpred.mutdb.org/</a> <a href="https://github.com/vpejaver/mutpred2">https://github.com/vpejaver/mutpred2</a>	[112]
LEAP	21	Private	[68]
LYRUS	4	<a href="https://github.com/jiaying2508/LYRUS">https://github.com/jiaying2508/LYRUS</a>	[113]

train LYRUS, and the VariBench testing dataset was used to assess its performance. Performance analysis revealed that LYRUS performed similarly to the most popular variant effect predictors. Six deep mutational scanning datasets for PTEN and TP53 were used to benchmark LYRUS' performance [113].

### Align Grantham Variation Grantham Deviation (Align-GVGD)

Align Grantham Variation Grantham Deviation (Align-GVGD) [114] is a freely available, web-based tool that uses the biophysical properties of amino acids and protein multiple sequence alignments to predict where missense variations in important genes will fall on a spectrum from enriched deleterious to enriched neutral. It classifies variants according to the level of cross-species conservation observed for a single missense substitution while considering the biophysical characteristics of the amino acids, and it's considered a non-ML method [114]. In the study by Tavtigian et al., an extension of the Grantham difference (A-GVGD) was used to classify missense variations in the BRCA1 gene. The method combined two techniques: the co-incidence of unclassified variants with clearly deleterious variants and the use of Grantham differences to analyze most missense variants. The

researchers used this approach to distinguish known neutral and deleterious missense variants into distinct sets and classified eight unclassified variants as neutral. This approach can be helpful in determining the functional impact of genetic variations in the BRCA1 gene, which is associated with an increased risk of BC and ovarian cancer [115].

### Sorting intolerant from tolerant (SIFT)

Sorting intolerant from tolerant (SIFT) [116] is a tool that predicts the deleteriousness of an amino acid substitution to a protein. SIFT was trained originally on lacI, lysosyme and HIV protease amino acid substitutions. It is frequently used to prioritize non-synonymous missense variants. An amino acid change may be tolerated, and the protein still functions normally, but sometimes, the protein might not tolerate a given amino acid change. SIFT categorizes the amino acid change as tolerated or deleterious to the protein's function. SIFT is categorized under the non-ML tools, as it considers protein conservation with homologous sequences alongside the severity of the amino acid change [116]. In multiple studies, SIFT has been shown to achieve high sensitivity levels in predicting the functional impact of variants in the BRCA1 and BRCA2 genes. In Poon [101], it was reported that SIFT

**Table 7:** Advantages and disadvantages of each tool.

Tool	Advantages	Disadvantages
CADD	<ul style="list-style-type: none"> <li>It supports systematic and objective labeling of variants.</li> <li>It can accommodate almost any feature tied to reference assembly coordinates.</li> <li>It has the capacity to score both coding and non-coding variants.</li> </ul>	<ul style="list-style-type: none"> <li>The label of the training dataset for any given variant provides a low estimate of whether the variant is benign or pathogenic.</li> </ul>
PolyPhen-2	<ul style="list-style-type: none"> <li>It has general robustness.</li> </ul>	<ul style="list-style-type: none"> <li>It has low specificity.</li> </ul>
Fathmm-MKL	<ul style="list-style-type: none"> <li>It can predict coding and non-coding regions.</li> </ul>	<ul style="list-style-type: none"> <li>There are limited non-coding datasets available.</li> </ul>
REVEL	<ul style="list-style-type: none"> <li>It is trained and tested on recently identified diseases and associated variants.</li> <li>It incorporates more individual predictors than prior ensemble methods.</li> <li>The training and testing sets used to train any component predictors have been removed to reduce overfitting.</li> </ul>	<ul style="list-style-type: none"> <li>The reliance on pathogenicity assertions from existing databases and predictors might be inaccurate and incomplete.</li> </ul>
CScape	<ul style="list-style-type: none"> <li>It can predict coding and non-coding regions.</li> <li>It is specifically developed for oncogenic variants.</li> </ul>	<ul style="list-style-type: none"> <li>There is a rareness of validated oncogenic variants in non-coding regions.</li> </ul>
DeepDriver	<ul style="list-style-type: none"> <li>It is cancer specific.</li> </ul>	<ul style="list-style-type: none"> <li>It was only trained on specific cancer types but not all.</li> </ul>
DrABC	<ul style="list-style-type: none"> <li>It is highly specific.</li> </ul>	<ul style="list-style-type: none"> <li>There are few carriers of GPs in CPGs other than BRCA1/2, and their endophenotypes are not well represented.</li> </ul>
RENOVO	<ul style="list-style-type: none"> <li>It relies on fewer features, so it is easier to recollect and apply for features of new variants.</li> </ul>	<ul style="list-style-type: none"> <li>The lack of gene- and disease-specific optimization gives uneven performance across variant classes.</li> </ul>
SVFX	<ul style="list-style-type: none"> <li>The CVD (Cardiovascular disease) cohort in the study had a unique strength of being a careful case-control study.</li> </ul>	<ul style="list-style-type: none"> <li>The lack of high-quality inversions and translocations in public databases limits its applicability to distinguishing disease-associated SVs from benign ones.</li> </ul>
Aljarf et al.	<ul style="list-style-type: none"> <li>It tailors gene-specific predictive methods to uncover variant-structure-function relationships.</li> </ul>	<ul style="list-style-type: none"> <li>The number of experimentally validated deleterious variants in BRCA1 and BRCA2 is limited.</li> <li>The training data are restricted to defined variants that are in protein regions identified to be involved with impaired DNA repair.</li> <li>The source code is not available.</li> </ul>
MutPred and MutPred2	<ul style="list-style-type: none"> <li>It has better performance over other pathogenicity predictors when information on biochemical, molecular or functional impact is available.</li> </ul>	<ul style="list-style-type: none"> <li>It is based on a small and biased dataset (CAGI dataset).</li> </ul>
LEAP	<ul style="list-style-type: none"> <li>It is usable and interpretable so can combine many different forms of evidence used for expert manual variant classification based on ACMG guidelines.</li> </ul>	<ul style="list-style-type: none"> <li>It needs a more mature database and further model tuning.</li> <li>A disease-specific model needs increased training data size and generalizability.</li> <li>The source code is not available.</li> </ul>
LYRUS	<ul style="list-style-type: none"> <li>Includes the variation number.</li> </ul>	<ul style="list-style-type: none"> <li>Dynamic-based features have a low impact score.</li> <li>It cannot be applied to proteins.</li> </ul>

**Table 8:** Overview of the data features used in each tool development.

Tool	Nucleotide-based	Amino acid-based	Protein-based	Conservation-based
CADD	✓	✓	✓	✓
PolyPhen-2	✓	✓	✓	✓
Fathmm-MKL	✓	✓	✓	✓
REVEL	–	✓	✓	✓
CScape	✓	✓	–	–
DeepDriver	✓	–	–	–
DrABC	✓	–	–	–
RENOVO	✓	–	–	–
SVFX	✓	–	–	✓
Aljarf et al.	✓	✓	✓	✓
MutPred and MutPred2	✓	✓	✓	✓
LEAP	✓	–	✓	✓
LYRUS	✓	✓	✓	✓

**Table 9:** Comparison of the performance of different pathogenicity prediction tools on BC data.

Tools	Relevance			Type (ML/Non-ML)	AUC	Reference
	Trained on conservation data	Trained on BC data	Tested on BC data			
CADD	✓	–	✓	ML	0.99	[117]
Polyphen	✓	–	✓	ML	0.88	
Gene-specific model	✓	✓	✓	ML	0.999	
SIFT	✓	–	✓	Non-ML	0.11	[113]
Polyphen	✓	–	✓	ML	0.64	
Lyrus	✓	✓	✓	ML	0.89	
SIFT	✓	–	✓	Non-ML	0.66	[71]
Polyphen	✓	–	✓	ML	0.628	
CADD	✓	–	✓	ML	0.621	
Revel	✓	–	✓	ML	0.63	[101]
SIFT	✓	–	✓	Non-ML	0.40	
Polyphen	✓	–	✓	ML	0.77	
SIFT	✓	–	✓	Non-ML	0.55	[118]
Polyphen	✓	–	✓	ML	0.87	
Revel	✓	–	✓	ML	0.97	
SIFT	✓	–	✓	Non-ML	0.85	

achieved 100% sensitivity in predicting BRCA1 and BRCA2 variants. Similarly, in Kerr *et al.* [100], it was reported that SIFT had 100% sensitivity in predicting both BRCA1 and BRCA2 variants. In Ernst *et al.* [99], it was also confirmed that SIFT had 100% sensitivity in predictions on both BRCA1 and BRCA2 variants. It's worth noting that high sensitivity does not imply high specificity, and it may have a high rate of false positives [99].

## Comparison

Many tools have been developed for pathogenicity prediction based on different algorithms and utilizing different training datasets. These differences give rise to slightly different results when comparing them with the same input dataset. The training datatypes and algorithms used to develop a given tool should be considered when selecting the most suitable tool for a given dataset. Table 5 summarizes the application types, the programming language and the algorithm used to develop each of the aforementioned tools along with the reference for each tool that can be referred to for any additional information needed. Table 6 shows the reference for the reliability and significance of each tool, along with the number of citations from Google Scholar. Table 7 shows the advantages and disadvantages of each tool. Moreover, Table 8 shows the functionality of each tool. Finally, Table 9 shows a comparison between the performance of ML-based tools and non-ML-based tools using the AUC values. The gene-specific model tool that is specialized for BC has shown higher AUC compared with other ML-based tools like polyphen-2 and CADD, which, in turn, have shown higher AUC compared with the non-ML based tool SIFT [117]. Similarly, Lyrus, which is another cancer-specific tool, has shown higher performance in terms of AUC compared with the other ML-based tool Polyphen and the non-ML-based tool SIFT [113]. Additionally, the ML-based tools Revel, CADD and Polyphen have shown higher AUC compared with SIFT when tested on the same dataset of BC variants [71]. The tools Polyphen, Revel and SIFT were also used in another study to assess their performance in predicting BC variants, and the ML-based tools have shown higher AUC compared with SIFT [101, 118].

The tools specifically developed and trained on BC data were the most accurate when testing for BC variants, followed by

cancer-specific and finally non-disease-specific tools. One of the most accurate tools discussed in this paper is the DrABC tool, as it was developed and trained on BC data. As only a limited number of tools were developed for specific targeted diseases like BC, developing new tools trained on detailed BC data or training existing tools on BC data mostly yields more accurate results for predicting BC pathogenic variants. As proven by several research including Mohammad and Borbala [117], Nikta *et al.* [76] and Hui-Heng *et al.* [72] that when the tool was developed for cancer in specific and was trained on either variants of a specific gene or a collection of variants from different genes, it has shown an enhanced performance compared with the tools developed for general purposes.

## CONCLUSIONS

With the rapid development of genomics and many successful genome projects, the known number of missense variants is increasing rapidly. Thus, it has become essential to learn more about the pathogenicity of such variants to predict, prevent or tailor the treatment for diseases. This review discusses several tools and databases that can be useful in predicting the pathogenicity of variants associated with BC. We provide an in-depth review of diverse databases that can be used, the types of variants included, the accessibility of the underlying data sources and the website of each database. We provide an example of the databases and tools used for prediction of BC pathogenicity. Among all the reviewed databases, we identify that the databases with cancer-specific genetic variants such as NCG, IntOGen and OncoKB are considered strong candidates for training BC-specific pathogenicity prediction tools. The rising issue that states that predisposition gene variants are inherited from humans themselves and not from other primates is not valid on the discussed tools, as none of the tools discussed was trained only on conservation data.

Moreover, a description of each tool, the type of application, the training data, the algorithm realized by the tool and the reliability and accessibility of the source code link were provided. The advantages and disadvantages of each of the discussed tools were provided to aid biomedical researchers in choosing the tool most suitable for a particular research project. The pathogenicity



prediction tools DrABC and CScape were shown to have outstanding performance in predicting BC pathogenic variants. We identify that the tools specialized by training on multiple diverse datasets from different databases for the same disease have shown higher accuracy and specificity, thereby helping clinicians in predicting and diagnosing BC as early as possible. The tools discussed in this review are not restricted to BC only; other cancers and sometimes other diseases pathogenic variants can be predicted using the same available tools. The same applies to the databases, in which they are inclusive of variants for several diseases and cancers, not only BC.

### Key Points

- Review genetic variant databases used specifically for the prediction of breast cancer pathogenicity.
- Review machine learning tools used for breast cancer variants pathogenicity prediction.
- Compare between different genetic variant databases and their influence of the prediction.
- Compare between different machine learning tools and their prediction performance using different genetic variant databases.

## FUNDING

This work was supported by the United Arab Emirates University through Strategic Research Program (No. Vot: 12R111) and Research Start-up Program (No. Vot: 12M109). This work is supported by ASPIRE Project VRI-20-10 (ADPMVRI) and funded by ASPIRE Abu Dhabi. R.M.A. is supported by a PhD fellowship from the United Arab Emirates University.

## DATA AVAILABILITY STATEMENT

All data generated or analyzed during this study are included in this published article.

## REFERENCES

1. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. *Neural Comput* 2006;**18**(7):1527–54.
2. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**(7553):436–44.
3. Shendure J, Balasubramanian S, Church GM, et al. DNA sequencing at 40: past, present and future. *Nature* 2017;**550**(7676):345–353.
4. National Cancer Institute. Cancer Stat Facts: Common Cancer Sites. 2022.
5. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019;**69**(1):7–34.
6. Spinelli A. Men account for a small fraction of breast cancer cases. Their fatality rate has soared compared with women's. 2019, Accessed on 15th September 2023, Retrieved from <https://www.statnews.com/2019/10/11/mortality-ra>.
7. Cao C, Liu F, Tan H, et al. Deep learning and its applications in biomedicine. *Genomics Proteomics Bioinformatics* 2018;**16**:17–32.
8. Omran NF, Abd-El Ghany SF, Saleh H, et al. Applying deep learning methods on time-series data for forecasting COVID-19 in Egypt, Kuwait, and Saudi Arabia. *Complexity* 2021;**2021**:6686745. <https://doi.org/10.1155/2021/6686745>.
9. El-Sappagh S, Abuhmed T, Alouffi B, et al. the role of medication data to enhance the prediction of Alzheimer's progression using machine learning. *Comput Intell Neurosci* 2021;**2021**:8439655.
10. Ford E, Carroll JA, Smith HE, et al. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;**23**(5):1007–15.
11. Sharma H, Rizvi MA. Prediction of heart disease using machine learning algorithms: a survey. *Int J Recent Innov Trends Comput Commun* 2017;**5**(8), 99–104. <https://doi.org/10.17762/ijritcc.v5i8.1175>.
12. Saleh H, Abd-El Ghany SF, Alyami H, Alosaimi W. Predicting breast cancer based on optimized deep learning approach. *Comput Intell Neurosci* 2022;**2022**:1820777.
13. Liu H, Chen Y, Li G, et al. Adaptive fuzzy synchronization of fractional-order chaotic (hyperchaotic) systems with input saturation and unknown parameters. *Complexity* 2017;**2017**:6853826.
14. Liu H, Pan Y, Li S, Chen Y. Synchronization for fractional-order neural networks with full/under-actuation using fractional-order sliding mode control. *Int J Mach Learn Cybern* 2018;**9**(7):1219–32.
15. Nirmala G, Kumar S. Deep Convolutional Neural Network for Breast Mass Classification from Mammogram. *Bioscience Biotechnology Research Communications*, 2020;**13**(13):203–208. <https://doi.org/10.21786/bbrc/13.13/28>.
16. Savage J, Ars E, Cotton RGH, et al. DNA variant databases improve test accuracy and phenotype prediction in Alport syndrome. *Pediatr Nephrol* 2014;**29**:971–7.
17. Ritter DI, Roychowdhury S, Roy A, et al. Somatic cancer variant curation and harmonization through consensus minimum variant level data. *Genome Med* 2016;**8**(1):117, 1–9.
18. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
19. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;**45**(D1):D777–83.
20. Landrum MJ, Lee JM, Benson M, et al. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018;**46**(D1):D1062–7.
21. Mottaz A, David FPA, Veuthey AL, Yip YL. Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics* 2010;**26**(6):851–2.
22. Li J, Shi L, Zhang K, et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res* 2018;**46**(D1):D1039–48.
23. Thangam M, Gopal RK. CRCDA - Comprehensive resources for cancer NGS data analysis. *Database* 2015;**2015**(1), 1–15.
24. Saunders CT, Baker D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J Mol Biol* 2002;**322**(4):891–901.
25. Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet* 2000;**16**(5):198–200.
26. Wang Z, Moult J. SNPs, protein structure, and disease. *Hum Mutat* 2001;**17**(4):263–70.
27. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9.
28. Ancien F, Pucci F, Godfroid M, Rooman M. Prediction and interpretation of deleterious coding variants in terms of protein structural stability. *Sci Rep* 2018;**8**:1–11.

29. Capriotti E, Altman RB. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics* 2011;**12**:1471-2105-12-S4-S3.
30. Richards S, Aziz N, Bale S, et al. 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* 2015;**17**(5):405-24.
31. Marinko JT, Huang H, Penn WD, et al. Folding and misfolding of human membrane proteins in health and disease: from single molecules to cellular proteostasis. *Chem Rev* 2019;**119**: 5537-606.
32. Niu B, Scott AD, Sengupta S, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* 2016;**48**(8):827-37.
33. Yip YL, Famiglietti M, Gos A, et al. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Hum Mutat* 2008;**29**:361-6.
34. Wright AF. Genetic Variation: Polymorphism and Mutation. *Encyclopedia of Life Sciences & 2005*, John Wiley & Sons, Ltd., <https://doi.org/10.1038/npg.els.0005005>.
35. Yue P, Li Z, Moul J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 2005;**353**(2): 459-73.
36. Dine J, Deng CX. Mouse models of BRCA1 and their application to breast cancer research. *Cancer Metastasis Rev* 2013;**32**(1-2): 25-37.
37. Deng CX. BRCA1: Cell cycle checkpoint, genetic instability, DNA damage response and cancer evolution. *Nucleic Acids Res* 2006;**34**(5):1416-26.
38. Sánchez H, Paul MW, Grosbart M, et al. Architectural plasticity of human BRCA2-RAD51 complexes in DNA break repair. *Nucleic Acids Res* 2017;**45**(8):4507-18.
39. Martinez JS, von Nicolai C, Kim T, et al. BRCA2 regulates DMC1-mediated recombination through the BRC repeats. *Proc Natl Acad Sci U S A* 2016;**113**(13):3515-20.
40. Harbeck N, Gnant M. Breast cancer. *Lancet* 2017;**389**(10074): 1134-50.
41. Ali R, Wendt MK. The paradoxical functions of EGFR during breast cancer progression. *Signal Transduct Target Ther*. 2017;**2**:16042-. <https://doi.org/10.1038/sigtrans.2016.42>.
42. Appert-Collin A, Hubert P, Crémel G, Bennasroune A. Role of ErbB receptors in cancer cell migration and invasion. *Front Pharmacol* 2015;**6**:283.
43. Jung MS, Russell AJ, Liu B, et al. A Myc activity signature predicts poor clinical outcomes in Myc-associated cancers. *Cancer Res* 2017;**77**(4):971-81.
44. Chen Y, Olopade OI. MYC in breast tumor progression. *Expert Rev Anticancer Ther* 2008;**8**:1689-98.
45. Pylayeva-Gupta Y, Grabocka E, Bar-Sagi D. RAS oncogenes: weaving a tumorigenic web. *Nat Rev Cancer* 2011;**11**: 761-74.
46. Hientz K, Mohr A, Bhakta-Guha D, Efferth T. The role of p53 in cancer drug resistance and targeted chemotherapy. *Oncotarget* 2017;**8**:8921-8946.
47. Varna M, Bousquet G, Plassa LF, et al. TP53 status and response to treatment in breast cancers. *J Biomed Biotechnol* 2011;**2011**:284584.
48. Roberts MR, Sucheston-Campbell LE, Zirpoli GR, et al. Single nucleotide variants in metastasis-related genes are associated with breast cancer risk, by lymph node involvement and estrogen receptor status, in women with European and African ancestry. *Mol Carcinog* 2017;**56**(3):1000-9.
49. Qu S, Long J, Cai Q, et al. Genetic polymorphisms of metastasis suppressor gene NME1 and breast cancer survival. *Clin Cancer Res* 2008;**14**(15):4787-93.
50. Lefebvre C, Bachelot T, Filleron T, et al. Mutational profile of metastatic breast cancers: a retrospective analysis. *PLoS Med* 2016;**13**(12):e1002201.
51. Cheng L, Zhou Z, Flesken-Nikitin A, et al. Rb inactivation accelerates neoplastic growth and substitutes for recurrent amplification of cIAP1, cIAP2 and Yap1 in sporadic mammary carcinoma associated with p53 deficiency. *Oncogene* 2010;**29**(42): 5700-11.
52. Loibl S, Darb-Esfahani S, Huober J, et al. Integrated analysis of PTEN and p4EBP1 protein expression as predictors for PCR in HER2-positive breast cancer. *Clin Cancer Res* 2016;**22**(11): 2675-83.
53. Hernandez-Aya LF, Gonzalez-Angulo AM. Targeting the phosphatidylinositol 3-kinase signaling pathway in breast cancer. *Oncologist* 2011;**16**(4):404-14.
54. Choi M, Kipps T, Kurzrock R. ATM mutations in cancer: therapeutic implications. *Mol Cancer Ther* 2016;**15**:1781-91.
55. Desmedt C, Zoppoli G, Gundem G, et al. Genomic characterization of primary invasive lobular breast cancer. *J Clin Oncol* 2016;**34**(16):1872-80.
56. Su Y, Wang X, Li J, et al. The clinicopathological significance and drug target potential of FHIT in breast cancer, a meta-analysis and literature review. *Drug Des Devel Ther* 2015;**9**:5439-45.
57. Berardi R, Morgese F, Onofri A, et al. Role of maspin in cancer. *Clin Transl Med* 2013;**2**(1):8.
58. Shahriar D, Mohammad A, Jahanbano S, et al. Maspin gene expression in invasive ductal carcinoma of breast. *Iran J Pathol* 2016;**11**(2):104-11.
59. Inoue K, Fry EA. Aberrant expression of cyclin D1 in cancer. *Sign Transduct Insights* 2015;**4**:1-13. <https://doi.org/10.4137/STI.S30306>. STI.S30306.
60. Nakagawa H, Fujita M. Whole genome sequencing analysis for cancer genomics and precision medicine. *Cancer Sci* 2018;**109**: 513-22.
61. Felicio PS, Grasel RS, Campacci N, et al. Whole-exome sequencing of non-BRCA1/BRCA2 mutation carrier cases at high-risk for hereditary breast/ovarian cancer. *Hum Mutat* 2021;**42**(3): 290-9.
62. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 2015;**1A**:A68-77.
63. Li MM, Datto M, Duncavage EJ, et al. 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 Diagn* 2017;**19**:4-23.
64. Horak P, Griffith M, Danos AM, et al. Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): joint recommendations of Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC), and Variant Interpretation for Cancer Consortium (VICC). *Genet Med* 2022;**24**(5):986-98.
65. Stenson PD, Mort M, Ball E, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum Genet* 2017;**136**: 665-77.
66. Cooper DN, Chen JM, Ball E, et al. Genes, mutations, and human inherited disease at the dawn of the age of personalized genomics. *Hum Mutat* 2010;**31**:631-55.

67. Stenson PD, Mort M, Ball E, et al. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Hum Genet* 2020;**139**:1197–207.
68. Lai C, Zimmer AD, O'Connor R, et al. LEAP: Using machine learning to support variant classification in a clinical setting. *Hum Mutat* 2020;**41**(6):1079–90.
69. Cline MS, Babbi G, Bonache S, et al. Assessment of blind predictions of the clinical significance of BRCA1 and BRCA2 variants. *Hum Mutat* 2019;**40**(9):1546–56.
70. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;**44**(D1):D862–8.
71. Yazar M, Ozbek P. Assessment of 13 in silico pathogenicity methods on cancer-related variants. *Comput Biol Med* 2022;**145**:105434.
72. Lin HH, Xu H, Hu H, et al. Predicting ovarian/breast cancer pathogenic risks of human BRCA1 gene variants of unknown significance. *Biomed Res Int* 2021;**2021**:6667201.
73. Nono AD, Chen K, Liu X. Comparison of different functional prediction scores using a gene-based permutation model for identifying cancer driver genes. *BMC Med Genomics* 2019;**12**(Suppl 1):22.
74. Shihab HA, Rogers MF, Gough J, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 2015;**31**(10):1536–43.
75. Grossman RL, Heath AP, Ferretti V, et al. Toward a shared vision for cancer genomic data. *N Engl J Med* 2016;**375**(12):1109–12.
76. Feizi N, Liu Q, Murphy L, Hu P. Computational prediction of the pathogenic status of cancer-specific somatic variants. *Front Genet* 2022;**12**:805656.
77. Gudmundsson S, Singer-Berk M, Watts NA, et al. Variant interpretation using population databases: lessons from gnomAD. *Hum Mutat* 2022;**43**:1012–30.
78. Lek M, Karczewski KJ, Minikel E, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 2016;**536**(7616):285–91.
79. Rofes P, del Valle J, Torres-Esquius S, et al. Bard1 pathogenic variants are associated with triple-negative breast cancer in a spanish hereditary breast and ovarian cancer cohort. *Genes (Basel)* 2021;**12**(2):1–11.
80. Syed AS, D'Antonio M, Ciccarelli FD. Network Of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Res* 2009;**38**(SUPPL.1):D670–5.
81. Repana D, Nulsen J, Dressler L, et al. The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens 06 Biological Sciences 0604 Genetics 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis 06 Biological Sciences 0601 Biochemistry and Cell Biology. *Genome Biol* 2019;**20**(1):1–12.
82. Hamosh A, Scott AF, Amberger JS, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**(DATABASE ISS):D514–7.
83. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, et al. IntOGen-mutations identifies cancer drivers across tumor types. *Nat Methods* 2013;**10**(11):1081–2.
84. Cerami E, Gao J, Dogrusoz U, et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012;**2**(5):401–4.
85. Borchert F, Mock A, Tomczak A, et al. Knowledge bases and software support for variant interpretation in precision oncology. *Brief Bioinform* 2021;**22**(6):1–17.
86. Chung IF, Chen CY, Su SC, et al. DriverDBv2: A database for human cancer driver gene research. *Nucleic Acids Res* 2016;**44**(D1):D975–9.
87. Gu H, Xu X, Qin P, Wang J. FI-Net: identification of cancer driver genes by using functional impact prediction neural network. *Front Genet* 2020;**10**:11:564839.
88. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* 2017;**2017**: 1–16: PO.17.00011. <https://doi.org/10.1200/PO.17.00011>.
89. Koeppel F, Muller E, Harlé A, et al. Standardisation of pathogenicity classification for somatic alterations in solid tumours and haematologic malignancies. *Eur J Cancer* 2021;**159**: 1–15.
90. Ng PKS, Li J, Jeong KJ, et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* 2018;**33**(3):450–462.e10.
91. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**(7391):603–7.
92. Stanfield Z, Coskun M, Koyutürk M. Drug response prediction as a link prediction problem. *Sci Rep* 2017;**9**(7):40321.
93. Gönen M, Alpaydin E. Multiple kernel learning algorithms. *J Mach Learn Res* 2011;**12**:2211–2268.
94. Ferroni P, Zanzotto FM, Scarpato N, et al. Risk assessment for venous thromboembolism in chemotherapy-treated ambulatory cancer patients. *Med Decis Making* 2017;**37**(2): 234–42.
95. Ferroni P, Roselli M, Zanzotto FM, Guadagni F. Artificial intelligence for cancer-associated thrombosis risk assessment. *Lancet Haematology* 2018;**5**:e391.
96. Won DG, Kim DW, Woo J, Lee K. 3Cnet: Pathogenicity prediction of human variants using multitask learning with evolutionary constraints. *Bioinformatics* 2021;**37**(24):4626–34.
97. Nakagomi H, Mochizuki H, Inoue M, et al. Combined annotation-dependent depletion score for BRCA1/2 variants in patients with breast and/or ovarian cancer. *Cancer Sci* 2018;**109**(2):453–61.
98. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Current Protocols in Human Genetics*, 2013;**76**:7.20.1–7.20.41. <https://doi.org/10.1002/0471142905.hg0720s76>.
99. Ernst C, Hahnen E, Engel C, et al. Performance of in silico prediction tools for the classification of rare BRCA1/2 missense variants in clinical diagnostics. *BMC Med Genomics* 2018;**11**(1):35. <https://doi.org/10.1186/s12920-018-0353-y>.
100. Kerr ID, Cox HC, Moyes K, et al. Assessment of in silico protein sequence analysis in the clinical classification of variants in cancer risk genes. *J Community Genet* 2017;**8**(2): 87–95.
101. Poon KS. In silico analysis of BRCA1 and BRCA2 missense variants and the relevance in molecular genetic testing. *Sci Rep* 2021;**11**(1):11114. <https://doi.org/10.1038/s41598-021-88586-w>.
102. Ioannidis NM, Rothstein JH, Pejaver V, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet* 2016;**99**(4):877–85.
103. Aljarf R, Shen M, Pires DE, Ascher DB. Understanding and predicting the functional consequences of missense mutations in BRCA1 and BRCA2. *Sci Rep* 2022;**12**(1):10458. <https://doi.org/10.1038/s41598-022-13508-3>.

104. Rogers MF, Shihab HA, Gaunt TR, Campbell C. CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci Rep* 2017;**7**(1):10458, 1–10. <https://doi.org/10.1038/s41598-017-11746-4>.
105. Rogers MF, Gaunt TR, Campbell C. CScape-somatic: distinguishing driver and passenger point mutations in the cancer genome. *Bioinformatics* 2020;**36**(12):3637–44.
106. Luo P, Ding Y, Lei X, Wu FX. DeepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front Genet* 2019;**10**(JAN):13, 1–12.
107. Liu J, Zhao H, Zheng Y, et al. DrABC: deep learning accurately predicts germline pathogenic mutation status in breast cancer patients based on phenotype data. *Genome Med* 2022;**14**(1):21, 1–15.
108. Favalli V, Tini G, Bonetti E, et al. Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm. *Am J Hum Genet* 2021;**108**(4): 682–95.
109. Kumar S, Harmanci A, Vythoeswaran J, Gerstein MB. SVFX: a machine learning framework to quantify the pathogenicity of structural variants. *Genome Biol* 2020;**21**(1): 274.
110. Weischenfeldt J, Symmons O, Spitz F, Korbel JO. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 2013;**14**:125–38.
111. Li B, Krishnan VG, Mort ME, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 2009;**25**(21):2744–50.
112. Pejaver V, Mooney SD, Radivojac P. Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Hum Mutat* 2017;**38**(9): 1092–108.
113. Lai J, Yang J, Gamsiz Uzun ED, et al. LYRUS: a machine learning model for predicting the pathogenicity of missense variants. *Bioinformatics. Advances* 2022;**2**(1):1–10.
114. Dorling L, Carvalho S, Allen J, et al. Breast cancer risks associated with missense variants in breast cancer susceptibility genes. *Genome Med* 2022;**14**(1):51, 1–17.
115. Tavtigian S, Deffenbaugh AM, Yin L, et al. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 2006;**43**(4):295–305.
116. Vaser R, Adusumalli S, Leng SN, et al. SIFT missense predictions for genomes. *Nat Protoc* 2016;**11**(1):1–9.
117. Khandakji MN, Mifsud B. Gene-specific machine learning model to predict the pathogenicity of BRCA2 variants. *Front Genet* 2022;13.
118. Gunning AC, Fryer V, Fasham J, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J Med Genet* 2021;**58**(8):547–55.