Chapter 1

Introduction

1. Introduction

1.1. Alzheimer's disease

Alzheimer's disease (AD), a neurodegenerative disorder, is characterised by dementia. Dementia involves progressive loss of comprehension, judgment, language, learning, memory and thinking skills due to a downturn in cerebral performance [1]. AD is the most common underlying cause of 50 – 75 % of dementia. Vascular dementia, frontotemporal dementia and dementia with Lewy bodies are the less common dementias [2]. According to a global projection, 135 million patients would have dementia by 2050, of which 71 million would be in the Asia-Pacific region. The region has displayed growth in the population of the elders, 80 years or above, from 0.8 % in 1990 to 1.4 % in 2012 with an estimated projection of 4.4 % by 2050. The report also suggests that India would be having about 12 million (9 %) people suffering from dementia by 2050, thus making the situation very challenging [3].

1.1.1. Symptoms

The noticeable early symptoms of AD include memory loss with difficulty in recalling things, i.e., recent events, visits or discussions. As the disease progresses, forgetfulness or memory lapse such as repetitive questioning, forget appointments, misplacing things, cannot find the way to a familiar place etc. becomes frequent. The patient is unable to focus, develops reduced thinking ability, poor multitasking performance, fall in reasoning ability and judgement and declined performance in the known and familiar tasks [4, 5]. The other neuro-psychiatric symptoms of the disease include apathy, irritability, mood swings, social withdrawal, aggression, suspicion on others, delusion, wandering, changes in the sleep cycle and insomnia etc. AD causes a high burden of suffering to patients, their families and caregivers [5].

1.1.2. Stages of Alzheimer's disease

1.1.2.1. Early-stage

The signs of the early stage of AD are often treated as the effect of old age by the patient, family and society.

- Unable to find the way that usually results in aimless wandering.
- Unable to handle finance and forgets to pay bills.
- Struggle associated with performing a routine daily task.
- Behavioural changes.
- Repetitive questioning.

1.1.2.2. Mid-stage

The symptoms get worsen as the disease progresses, with damage extending to other parts of the brain.

- Memory loss progresses with the patient being unable to recognise family, friends and relatives.
- Unable to learn a new task.
- Difficulty in performing multi-step tasks such as cooking, getting dressed.
- Waning sense of smell and taste due to damage to the sensory processing region of the brain.
- Language, reasoning and thought impairment.

1.1.2.3. Late-stage

The later stages involve severe damage to the brain and a reduction in its size and volume.

- Unable to communicate.
- Completely dependent on the caregiver.

- Bedridden.

1.1.3. Diagnosis of Alzheimer's disease

The Alzheimer's Association suggests a few guidelines for the diagnosis of AD. The diagnostic test involves physical examination, followed by neurological examination such as testing the body reflexes, muscle coordination and strength, sensation and eyeball movement. The further testing is performed by a mental examination, viz. Mini-Cog test and mini-mental state exam (MSME). A Mini-Cog test involves completion of the two tasks:

1. A patient is made to remember the name of three objects, and after a certain time, it is asked to repeat the name.
2. In the second test, the patient is made to draw a clock with all 12 numbers and asked to indicate a time specified.

The results of the examination indicate that whether there is need for a further investigation or not.

In MSME, a questionnaire is asked by a health practitioner to assess a variety of ordinary mental skills. The maximum score of MMSE is 30 points. In case of mild dementia, a score of 20 to 24 is expected. Moderate dementia is indicated by a score of 13 to 20, while severe dementia by a score of fewer than 12. A person with AD would have average decrease in MMSE score of two to four points every year [6].

Depression and mood assessment tests are also carried out to diagnose the other symptoms present in AD. A brain imaging through magnetic resonance imaging (MRI) or computed tomography (CT) is recommended to further access the underlying cause of dementia. In some instances, dementia may be due to tumour, stroke or trauma but not AD. The florbetapir, florbetaben and flutemetamol are amyloid-specific PET ligands that are employed for the diagnosis of AD pathology [7-9]. The other tests also involve estimation of Aβ_{42}, hyperphosphorylated tau (p-tau), and total tau protein in cerebrospinal fluid [10].

1.1.4. Pathology associated with Alzheimer's disease

Dr. Alois Alzheimer published the findings of an autopsy on Auguste Deter, a 55-year-old lady who died of a degenerative behavioural and cognitive illness in 1907. Deter's brain had two distinct characteristics, i.e., plaque and neurofibrillary tangle (NFT) made up of amyloid-β and hyperphosphorylated tau protein, respectively. Besides the two classical hallmarks, AD is associated with complex pathophysiological factors that are not completely revealed yet. A plethora of pathways and targets are involved in AD.

1.1.4.1. Amyloid-β hypothesis

The extracellular amyloid-β (Aβ) plaque consists of numerous variants of amyloid protein having a molecular weight of about 4 KDa with diverse C-terminals [11]. The variants of range A$\beta_{1\text{-}40}$ and A$\beta_{1\text{-}39}$ are soluble forms, while the A$\beta_{1\text{-}42}$ are susceptible to misfolding and aggregation [11]. The presence of Aβ in the grey matter of the brain is positively correlated with the severity of dementia. Amyloid precursor proteins (APPs) are membrane-bound proteins belonging to the amyloid precursor-like proteins family. Eight transcripts of APP are produced by alternative splicing. The protein contains 695 amino acid residues, is predominantly found in the central nervous system (CNS), and the other 751 and 770 amino acid proteins are ubiquitously expressed [12]. APP appears to modulate cell growth and survival, motility, neurite outgrowth. APP ectodomain helps in cognitive function and synaptic density [13]. Further, APP ectodomain also acts on Death Receptor 6 that activates caspase 3 and 6, produces axonal and cell body apoptotic degeneration responsible for axonal degeneration [14].

The APP is processed by a particular class of enzymes known as secretase. The normal catabolism involves, cleavage of APP by α-secretase, which could be subsequently cleaved by β- and γ-secretases [15]. Further, α-secretase produces soluble ectodomains(sAPPα) and various subunits viz. presenilin 1 (PS1) or presenilin 2 (PS2), type I transmembrane glycoprotein, nicastrin (Nct) and Aph-1 and Pen-2, two transmembrane proteins, make up the γ-secretase. During cleavage by a γ-secretase, a p3 subunit is formed and released in extracellular space (non-amyloidogenic pathway) [16]. In contrast, the action of the β-secretase produces soluble ectodomains(sAPPβ) and subsequent cleavage by γ-secretase results in Aβ fragments (amyloidogenic pathway). The deposited $A\beta_{1\text{-}42}$ oligomers cause increased tau hyperphosphorylation, oxidative injury resulting in loss of synapses and mitochondrial toxicity [17]. It also affects the activity of glycogen synthase kinase-3 (GSK-3), leading to neurodegeneration. Similarly, the hippocampal injection of the Aβ in the mice also resulted in amnesia [18, 19].

1.1.4.2. Tau hypothesis

The tau protein is responsible for the assembly of microtubules of the cytoskeleton. The monomeric tau protein has a molecular weight of about 55–74 KDa with six variants resulting from various splicing modes and has three or four microtubule-binding repeats [20]. Its dimeric form is made up of two monomeric units connected in anti-parallel fashion through disulfide bonds. Further, trimers, small soluble oligomer, insoluble granular tau oligomer and filaments are the various forms of tau protein [21]. NFTs are the proteaceous bodies composed of the aggregates of hyperphosphorylated microtubule-associated protein tau [22]. The origin of NFTs is found in the entorhinal cortex, subcortical regions and trans-entorhinal area. It progresses toward the hippocampal region and neocortex. The presence of NFTs in these regions is usually termed as Braak stages and shows a good correlation with the disease progression [23]. Various proteolytic event by enzymes viz. pruning of tau protein at Glu391 or at Asp421 by caspase promotes its aggregation [24-26]. The deamination of Asp or Glu amino acid residues, cleavage mediated through thrombin or truncation at C-terminal are other contributing factors [27-

30]. GSK-3β, Mitogen-activated Protein/ERK Kinase Kinases (MEKK) and c-Jun N-terminal kinase 3 (JNK3) have shown involvement in tau aggregation in an *in vitro* study [31]. Several studies indicated that phosphorylation of tau protein at specific sites viz. Thr321, Ser296, Ser404 and Ser422 led to aggregation [28, 32]. Human post-mortem studies showed a strong correlation between the severity of dementia and neurodegeneration with the presence of NFTs [23, 33-37].

GSK-3β is serine/threonine kinase and is made up of two domains: β-strand domain present at N-terminal (25-138) and a C-terminal α-helical domain formed by residues 139-343 [38]. The adenosine triphosphate (ATP) binding pocket is present at the interface of both the domains and the residues 200 – 226 forms the activation loop. The phosphorylation of the substrate occurs at a SXXXS motif [39]. It controls various cellular processes and acts as a master switch to regulate several signalling pathways [40, 41]. GSK-3β is associated with tau hyperphosphorylation, increased Aβ production, memory loss and inflammatory reactions in AD [42]. It also downregulates the synthesis of acetylcholine (ACh) and is involved in apoptosis and neuronal death [43, 44]. It is over-expressed in the hippocampal region of AD patients [45].

1.1.4.3. Cholinergic hypothesis

The cholinergic neural pathway in CNS plays an important role in cognition and behaviour [46]. It is observed that the cholinergic input to the cerebral cortex is disrupted in AD, that results in a lack of attention and decision-making process [47]. Further, the administration of scopolamine, a competitive antagonist of acetylcholine at muscarinic receptors by causing blockade of cholinergic receptors in the CA3 region of the hippocampus of the rats, showed impaired information encoding and memory formation [48]. The administration of scopolamine impaired memory storage and retrieval along with cognitive non-memory tasks in normal young human volunteers while protecting immediate memory [49]. In AD patients, the severity of the memory loss and neurodegeneration was positively correlated with the damage to the hippocampal region through synaptic alterations [50]. The reduction in choline acetyltransferase (ChAT), an ACh synthesising enzyme, activity was observed in the hippocampus and other regions of the brain [51, 52]. Defects in ACh synthesis could underpin cognitive failure and infer through a direct correlation observed between the ChAT activity and mental test score in dementia patients [53]. The cholinergic tract present in the nucleus basalis of Meynert is severely damaged with the disease progression and loss of ChAT activity in the patients [54]. The loss of cholinergic neurons also led to subsequent damage to dopaminergic pathways. The cholinergic neurons are projected to control the laterodorsal tegmental nucleus and pedunculopontine nucleus that led to nucleus accumbens. It resulted in neuropsychiatric symptoms such as including apathy and depression [55-57].

1.1.4.4. Excitotoxicity and NMDA receptors

Glutamate is an excitatory neurotransmitter present in almost all the synapses of the brain. In the resting state, the synaptic concentration of the glutamate is 0.6 µM, which rises to 10 µM during signal transduction in the synapse [58, 59]. Glutamate acts through 'ligand-gated ionotropic glutamate receptors' (iGluRs) [60]. N-methyl-d-aspartate (NMDA) receptors are one of the voltage-gated iGluRs, which are activated by the removal of Mg^{+2} ion that causes receptor blockade during resting conditions [61]. NMDA receptor on activation causes a rapid influx of Ca^{+2} ion, but have slow ligand-gated kinetics. The prolonged-release of glutamate, in considerable concentration, results in long term potentiation through the strong activation of NMDA receptor that activates a Ca^{+2}/calmodulin dependent protein kinase II (CaMKII) and increases synaptic strength. In the case of modest activation of these receptors, activation of phosphatases causes long term depression [62]. Increased glutamatergic signalling causes excitotoxicity leading to nerve cell damage or death [63]. The modest and prolonged activation of NMDA receptors on post-synaptic membrane in AD causes sustained influx of Ca^{+2}, leading to gradual synaptic loss and cell death. The excitatory amino acid transporter 2 (EAAT2), present on neighbouring astrocytes involved in glutamate reuptake, is also impaired in AD and cause prolong availability of the neurotransmitter contributing towards excitotoxicity [64]. Further, Aβ causes modulation of NMDA receptors and elevates its synaptic current leading to excitotoxicity [65]. The activation of NMDA receptors also requires D-serine or glycine as a co-agonist. In AD, D-serine and serine racemase expression are increased and contribute toward excitotoxicity [66, 67].

1.1.4.5. Oxidative stress

Oxidative stress is defined as an imbalance in the redox component that leads to increased production of reactive oxygen species (ROS). Usually, the molecular oxygen is utilised by the cytochrome enzymes of the mitochondrial respiratory chain and the residual amount is converted to superoxide ($O_2^{\bullet}$) radical and hydrogen peroxide. However, the excessive production of these ROS causes damage to the tissues mediated by further production of hydroxyl radical ($OH^{\bullet}$) and other oxidants. Various enzymes such as superoxide dismutase (SOD), cytochrome c, glutathione peroxidase, ubiquinol etc., are responsible for neutralising ROS [68].

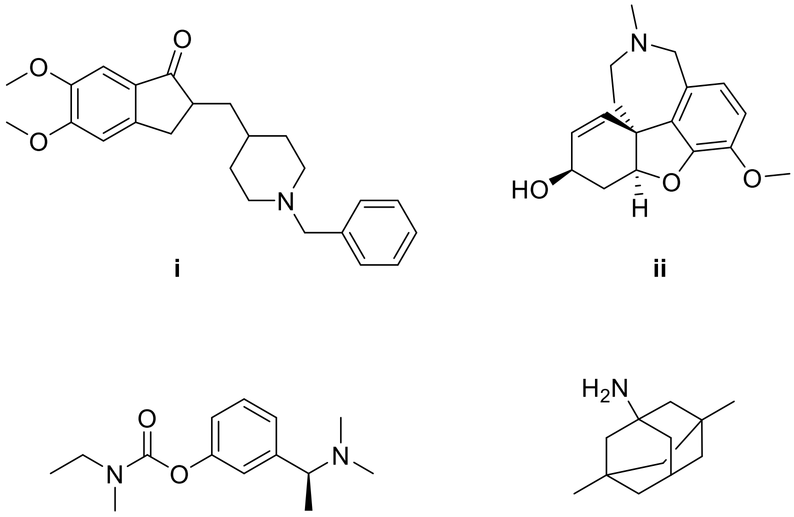
The N-terminal of Aβ has a metal-binding domain and is the site for binding of zinc and copper ions. These ions are responsible for converting superoxide and hydrogen peroxide to hydroxyl radical and lead to oxidative damage [69]. Further, the polyunsaturated phospholipids found in the plasma membrane of the neurons undergo lipid peroxidation causing brain damage [70]. The increased oxidative stress also causes damage to DNA and DNA-protein cross-linking [71].

1.1.5. Management of Alzheimer's disease

The management of cognitive symptoms of AD relies on four small molecules, three cholinesterase (ChE) inhibitors (donepezil, galantamine, and rivastigmine) and one NMAD receptor antagonist (memantine) (**Figure 1.1**), that are approved for use in the European Union, India and United States. More than a decade ago, memantine was the last approved small molecule for the treatment of AD [72].

The cognitive defect and memory loss is usually associated with the loss of cholinergic neurons. The increased concentration of ACh through inhibition of acetylcholinesterase (AChE) by the inhibitors in the synaptic cleft cause improvement in the symptoms. Donepezil (**i**) is a piperidine based reversible mixed AChE inhibitor with high selectivity toward AChE with IC_{50} value of 22 nM, against human erythrocyte AChE. It has low peripheral side effects at a dose of 5 – 10 mg per day with low oral bioavailability of 43 %, good blood-brain permeability and an extensive distribution volume (14 L/kg) [73, 74]. It is metabolised by CYP2D6 and CYP3A4 enzymes. Donepezil interacts with the amino acid residues of the peripheral anionic site (PAS) and anionic site (AS) of the enzyme [75]. Galantamine (**ii**) is a tertiary phenanthrene alkaloid obtained from *Galanthus woronow* and other plants of *Amaryllidaceae* family. Galantamine causes reversal of the tubocurare-induced muscular paralysis [76]. It interacts with AS and acyl binding pocket of the enzyme and is more selective towards AChE than butyrylcholinesterase (BChE) (1252: 1). The recommended dose is 8 – 32 mg/day with good oral absorption 85 – 100 % and C_{max} of 84.7 µg/ml reaching in 0.5 – 2 hours. It is also metabolised by CYP2D6 and CYP3A4 [77].

Rivastigmine (**iii**) shares the structural analogy of physostigmine and is a carbamate. It is a pseudo-reversible inhibitor that binds with esteratic and anionic sites with higher selectivity towards AChE present in CNS [78]. The recommended therapeutic dose is 6 – 12 mg/day. It displays an excellent oral absorption of more than 90%, with C_{max} reaching in 1 hour. It is rapidly metabolised by AChE by decarbmylation. The metabolite *(-)-S-3-[1-dimethyl)ethyl]phenol* is 10-folds more potent than the drug [79]. However, it is a dual inhibitor of AChE as well as BChE. Memantine (**iv**) is a *3,5-dimethyladamantan-1-amine* which is a non-competitive NMDA receptor antagonist with good bioavailability and C_{max} achieved in 3 – 8 hours. It has a large volume of distribution of about 9 – 11 L/kg. It is an 'uncompetitive low-affinity voltage-dependent antagonist' that binds to the NMDA receptor with a higher affinity than Mg^{2+} ion [80]. The recommended maximum daily dose of the drug is 20 mg/day. Aducanumab is an IgG1 anti-Aβ antibody that targets Aβ aggregate and has been recently approved by US food and drug administration [81].



i **ii** **iii** **iv**

Figure 1.1 Chemical structure for FDA approved drugs for Alzheimer's disease

1.2. Structure-based drug design

Structure-based drug design (SBDD) is a rational drug design technique that involves the use of the three-dimensional (3D) structure of the protein molecule for the identification of novel ligands [82]. Structure-based virtual screening (SBVS), molecular docking, and molecular dynamics (MD) simulations are the prevalent computational techniques employed in SBDD. The first approach involves the search of ligands from the large database that fit into the active site of the target protein by using a docking program. The other approach involved is *de novo* design with the stepwise building of a molecule using various smaller fragments within the binding site using the cavity constrain [83]. The third approach involves optimising the obtained ligands by testing potential analogues within the binding cavity [84].

1.2.1. Molecular docking

Molecular docking is the process of identification of the preferred orientation of a ligand in the cavity of the protein. The preferred orientation of the ligand is used to calculate binding affinity with the macromolecule [85]. It is one of the methods used to perform virtual screening. Molecular docking identifies the shape complementarity between the protein-ligand, which involves the use of surface area descriptors. It is a quick method, but does not take into account molecular flexibility and movements [86]. The other method uses the simulation principle and treating protein and ligand as flexible entities, usually referred to as flexible docking [87]. It could also treat protein as a rigid body and complete flexibility is provided to the ligand for rotational and translational changes [88]. The conformational spaces are vast for the protein and ligand. Usually, flexible docking involves the exhaustive search of the ligand conformational space. At the same time, selective conformations are used for protein to generate the protein-ligand complex. The conformational search could be carried out as a systematic or stochastic search. It could be done through MD simulation too. The genetic algorithm that identifies low potential energy poses to select individuals are subjected to further iterations [89]. The generated poses are then evaluated by using a mathematical function called the scoring function (SF). The classical SF is based on molecular mechanics to calculate the free binding energies of ligand conformation with binding site residues. It is usually represented as:

$$\Delta G_{binding} = \Delta G_{internal} + \Delta G_{vibrational} + \Delta G_{torsional} + \Delta G_{conf} + \Delta G_{solvent}$$

where Δ G is Gibbs's free energy.

However, the other SF is also gaining popularity viz., knowledge and machine learning (ML) based SF. The empirical SF is an interaction-based SF that involves the use of protein-ligand contact and the associated surface area for calculations. In the case of the knowledge-based SF, a statistical "potentials of mean force" is derived using known protein-ligand interactions with an assumption that some interactions are more favourable than others [90]. ML-based SF is gaining popularity due to the problem-specific approach and high prediction accuracy. These SF focuses on the protein-ligand interaction data to draw inferences and outperforms classical SF. They are generally customised or tailor-made SF for a specific macromolecular target [91].

1.2.2. Homology modelling

Homology modelling is a protein structure prediction technique based on the notion that two proteins with high sequence similarity share structural similarity. Hence, the 3D structure could serve as a template for predicting the structure of the other protein. It involves various steps:

- Template sequence search through Basic Local Alignment Search Tool (BLAST).
- Sequence alignment.
- Corrections of alignment to confirm the conserved residues are aligned correctly among the target-template sequences.
- Generation of the protein backbone.
- Modelling of the loops.
- Modelling of the side chains using rotamer libraries.
- Model optimization through energy minimisation.
- Validation of the developed model.

Various web servers are available for homology modelling, such as Swiss Model, ESyPred3D, Modeller, and Yasara. The model validation is usually carried out by stereochemical evaluation of the amino acid residues in the Ramachandran plot [92]. A Ramachandran plot indicates the relationship between two torsion angles viz. phi (φ)and psi (ψ) of the residues present in a protein [93]. The average values are φ= -57˚ and ψ= -47˚ for an α-helix and φ= -130˚ and ψ= 140˚ for a β-sheet. The residues centred around these torsional angle values are in the favourable region of the plot. The good quality of the model should have low outlier residues [94].

1.2.3. Molecular dynamics

MD is a method of exploring the evolution of a system by calculating the movement of the particles over a specific time period. The trajectories of the particles, atoms and/or molecules are calculated by solving Newton's equations of motion and the potential energies of particles and the forces between them are usually calculated through interatomic potentials or molecular mechanics force fields [95]. It involves four steps:

1. Initialize the system with total momentum as zero.
2. Determine the forces acting on each particle.
3. Integrate Newton's equation of motion.
4. Steps 2 and 3 are repeated for the desired time.

MD helps in understanding the folding and unfolding of a synthesised protein. It assists in the exploration of conformational changes in the protein with respect to time as well as other conditions, i.e., pH, the molarity of a solution, temperature etc. The catalytic mechanism, including the changes in the active site, substrate recognition, the transition state during catalysis, could also be explored by using MD [96].

1.3. Ligand-based drug design

Ligand-based drug design (LDBB) involves the use of a series of known inhibitors or antagonists or substrates of a protein for the identification of new molecules. It is generally used in the absence of 3D structure of the protein.

1.3.1. Pharmacophore modelling

A pharmacophore is an electronic or steric feature of a molecule responsible for interaction with the target protein [97]. A pharmacophore model helps to understand how different ligands can bind to a similar binding site of the receptor/enzyme. Hydrogen bond acceptor, hydrogen bond donor, aromatic ring, anion and cation are the five primary features used for the development of the pharmacophore model [98]. A pharmacophore is built by alignment and superimposition of various active ligands and is usually referred as a ligand-based pharmacophore model. On the other hand, a structure-based pharmacophore uses a protein-ligand complex or 3D structure of the protein to identify the interacting feature with the binding site [99].

1.3.2. Quantitative structure-activity relationship

Quantitative structure-activity relationship (QSAR) is a computational method for determining the relationship between chemical structures and a chemical property or biological activity. QSAR is based on the assumption that compounds with similar structural or physicochemical features have similar activity. The physicochemical characteristics of active molecules and biological activity are linked in a quantifiable way [100]. The statistical techniques that are commonly used for the development of QSAR models are multivariable linear regression analysis, partial least square and principal component analysis. QSAR is used to identify the essential features responsible for the activity [101].

1.4. Machine learning

ML is a type of data analysis that involves the development of analytical models with an intrinsic ability to analyse the data, identify the patterns and make judgments with little or no human aid. It is a subdomain of artificial intelligence. ML involves the use of various domains of knowledge viz. statistics, data mining, data visualisation and coding.

1.4.1. Approaches to machine learning

1.4.1.1. Supervised learning

Supervised learning is a ML technique in which models are trained by using well-labelled training data and then predict the output based on that data. The loss function is used to assess the algorithm's correctness, and it is adjusted until the error is suitably minimised. Supervised ML is of two types. Classification is used to classify a dataset and assign them to a specific class. It recognises various features in the dataset to conclude the labelling of a data instance. Support vector machines (SVM), k-nearest neighbour (KNN), decision trees and random forest (RF) etc., are classification algorithms. Regression is used to identify the dependency of a predictor variable on independent input variables. Linear regression, logistical regression, ridge regression and polynomial regression are regression algorithms.

1.4.1.2. Unsupervised learning

The unsupervised learning uses a collection of unlabelled data and detects the structure such as groups or clusters in the dataset. It identifies the common features or patterns in the data and the feedback depends on the presence or absence of the commonalities in each new data instance used for training.

1.4.1.3. Semi-supervised learning

In this technique, unlabelled data are combined with a modest amount of labelled data which significantly enhance learning accuracy. The learning fall between supervised and unsupervised learning.

1.4.2. Machine learning algorithms

A ML algorithm is a technique performed on the data to develop a model. It is a computer program that adapts to new data and improves its performance. Logistic regression is a classification technique that uses the coefficients to combine the input variables in a linear fashion and calculates the likelihood of a specific event or class occurring. The predictions are converted to a sigmoid function, and odds are calculated for each class to produce one of the binary outputs. SVM classifier identifies a manifold hyperplane to distinguish two or more classes in an n-dimensional feature space. To allow flexibility of the hyperplane, an important parameter termed slack variable (C) is introduced, which frequently results in a generalised model with a soft margin. Typically, the hyperplane is transformed using linear algebra, i.e., the linear kernel. Polynomial and radial functions are also employed as SVM kernels as well. It is a successful technique, especially in higher dimensional feature space, as it makes decisions using only a subset of training points called the support vector [102]. The KNN is a classification algorithm that employs the concept of searching all the closest neighbours around a new unknown data point. The labels of nearby points are allocated to the new data point. Although the Euclidean distance is most usually utilised, other popular distances such as Hamming, Manhattan, and Makowski can also be used [103]. The Bayes theorem is used to create a naive Bayesian categorization. It is assumed that for a particular class, the individual features utilised for prediction are conditionally independent of one another. Gaussian naïve Bayes assumes that the likelihood of features has gaussian distribution [104]. A perceptron classifier is an artificial neural network that takes in data and processes it using weighted values before transforming it with an activation function to produce output in the form of a categorical variable or probability of a class. Perceptron also iterates to reduce error by modifying the weights [105]. Ensemble algorithms integrate the predictions produced by a group of predictors developed on the dataset to derive the final label. A decision tree is a tree-like structure in which each node represents a feature test, the branch represents an outcome, and the leaf provides a class label. The RF classifier uses the bootstrap aggregation technique, but at each node a random set of features is utilised to determine the next decision. Adaptive boosting, also known as adaboost classifier, uses decision trees to construct a sequential model, with each new model correcting the error of the preceding one while giving properly predicted data a larger weight [106].

Linear regression algorithm uses a linear combination of features to determine the relationship between dependent and independent variables. The ordinary least squares method typically minimises the cost function, i.e., residual sum of squares between observed and predicted target values. A cost function calculates the distance between the predicted and expected output values. Besides the residual sum of squares, the ridge regression has a modified cost function with an additional L2 penalty, i.e., summation of square of the magnitude of coefficients. The lasso regression has a further additional L1 penalty, just the sum of the coefficients along with the residual sum of squares [107]. In addition to the residual sum of squares between observed and anticipated targets, elastic net regression utilises both L1 and L2 penalties in the cost function [108]. Support vector regression (SVR) identifies a hyperplane and the surrounding decision boundaries in which most training set points fall [109]. RF regression is similar to extra-tree regression. It makes a prediction using a forest of various decision trees [110]. Bayesian ridge regression develops a model by iteratively maximising the marginal log-likelihood of the observations. Huber regression uses the Huber loss, an absolute error that becomes quadratic when the difference between observed and predicted target values is minimal. Except for the linear activation function and mean squared error as the loss function, MLP regression is comparable to classification [111].

1.4.3. Validation of machine learning models

1.4.3.1. Classification

The developed models are validated by using the confusion matrix using accuracy, precision, recall and F1-score on the independent test dataset [112].

$$Accuracy\ = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+FN+TN}}$$

$$Precision\ = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$$

$$Recall\ = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$$

$$F1\,score\ = \frac{2*Precison*Recall}{Precision+Recall}$$

where TP: true positive, TN: true negative, FP: false positive and FN: false negative

1.4.3.2. Regression

The coefficient of determination (r^2) defines the dependence of one variable on another and ranges between 0 to 1. A higher value of r^2 represents a better fitting of the line or manifold plain on the data.

$$R^2 = \frac{SSR}{SST}$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SST = \sum_i (y_i - \bar{y})^2$$

where, y_i, $\hat{y}_i$– observed and predicted value of i^{th} observation. $\bar{y}$ is the mean value of the observation in the dataset.

Mean absolute error (MAE) and mean squared error (MSE) are other essential parameters and should be as low as possible.

$$MAE = \Sigma|Yobs - Ypred|$$

$$MSE = \Sigma(Yobs - Ypred)^2$$

The Q^2ext-based metrics, i.e., Q^2_{F1} and Q^2_{F2} should be greater than 0.5. The lower value indicates that the model fits better on the training set but have poor predictivity on an independent test dataset, i.e., overfitting.

$$Q^2F_1 = \frac{\Sigma(Y_{obs(test)} - Y_{pred(test)})^2}{\Sigma(Y_{obs(test)} - \bar{Y}_{(training)})^2}$$

$$Q^2F_2 = \frac{\Sigma(Y_{obs(test)} - Y_{pred(test)})^2}{\Sigma(Y_{obs(test)} - \bar{Y}_{(test)})^2}$$

where Yobs(test) and Ypred(test) are the observed and predicted values of the test compounds. $\bar{Y}_{(training)}$ and $\bar{Y}_{(test)}$ are the mean values of the observed activity of the training and test sets, respectively.