Chapter 1

Introduction

1.1 Main Problem

T-cells are important mechanisms within the human body that establish and maintain immune system responses, homeostasis, and memory. It expresses a receptor that identifies specific foreign antigens from pathogens, tumors, and the environment. T cells also have an immunological memory capability and are part of the body's machinery to create self-tolerance (Kumar et al., 2018). This machinery is vital in the adaptive immune system, recognizing the antigens belonging to foreign pathogens or aberrant forms of internal peptides (Gielis et al., 2019). Foreign antigens are expressed by the major histocompatibility complex (MHC) on the surface of nucleated host cells or antigen-presenting cells. Interactions between the t-cell receptors, or TCR, would induce a cascade of signaling pathways to initiate an immune system response specific to the identified antigens (Shah et al., 2021).

The relevancy of T-cells in the human immune system against cancer, autoimmune diseases, and pathogenic infections highlights these cells' significant importance in the immune system. Dysfunction in the mechanism of these t-cell responses poses a serious issue that could be categorically described as systemic autoimmunity (Schwartz et al., 2020). Several examples of t-cell mediated autoimmunity are atherosclerotic cardiovascular diseases (CVD), rheumatoid arthritis (RA), myositis, psoriasis, systemic lupus erythematosus (SLE), and vasculitis (Bluestone et al., 2015; Schwartz et al., 2020). For a T-cell immune system to activate, a complementary epitope should be recognized by the TCR and soluble antibodies (Sidney et al., 2020).

The method by which a T-cell receptor (TCR) recognizes an epitope is very sensitive to several factors that are present in the environment (Sidney et al., 2020). A different host, epitope organism source, immune system reaction pathway, or method of response quantification would give rise to a different

cascade of immune system responses. The importance of classifying epitopes can be considered paramount since it has a wide-ranging application in various areas such as therapeutics (Wilson & Andrews, 2012), diagnostics (Ahmad et al., 2016), and peptide-based vaccines (Ahmad et al., 2016; L. Dudek et al., 2010).

1.2 Specific Problem

Epitope mapping with peptides is the most extensive method of determining the epitope or peptide against an antibody (Bosshard, 1995). However, this method is both highly time-prohibitive and cost-inefficient (Steele et al., 2006). Other procedures using mass spectrometry have been introduced that provide better performance. Epitope peptide bound to tumor antigen has been successfully identified using a combination of high-performance liquid chromatography and mass spectrometry (Haslinger et al., 2006). Due to the above reasons, alternative TCR-epitope determination methods are essential to moving the field forward. One of the possible approaches is by using artificial intelligence (AI) based approaches.

Since data on protein sequences can be obtained through laboratory techniques, a computational approach to predicting epitope binding has been developed using AI approaches (Gielis et al., 2019; Isabell Jurtz et al., 2018; Isacchini et al., 2021; Jokinen et al., 2021; Luu et al., 2021; Montemurro et al., 2021; Moris et al., 2021; Sidhom et al., 2021; Springer et al., 2020; Tong et al., 2020). It has also evolved to become a crucial part of developing effective cancer immunotherapies (Singh-Jasuja et al., 2004). Three main databases—McPAS-TCR, VDJdb, and Immune Epitope Database (IEDB), each of which includes CDR3α and CDR3β information—store the majority of the data that are available for public use (Bagaev et al., 2020; Tickotsky et al., 2017; Vita et al., 2019). Many models have been developed to predict the TCR-epitope binding using data from individual databases or in combination with one another.

1.3 Proposed Method

In this work, we propose a deep learning architecture tailored to the previous problem space as part of our study. First, for handling direct amino acids sequence embedding, we utilize statistical mechanisms to extract specific properties since, in published work, it effectively improves the accuracy of the classification and can enrich feature sparseness (T. Liu et al., 2020; G. Wang et al., 2020). Second, to optimize the classifier's performance, we leverage transformer architecture that uses an attention mechanism that is broadly known for handling problems in LSTM and RNN (Recurrent Neural Network) model. This mechanism will directly lead the model to learn and focus on exact information from extracted features instead of using other models, such as LSTM, to extract the amino acid sequence. In addition, transformers offer a more efficient model architecture due to the parallel computation capability (Wen et al., 2022). It also enables using smaller layers in the architecture while achieving better performance. However, transformers' main benefit is the ability to interpret long-range feature dependencies and interactions in structured data such as epitope classification. An alternative fusion model with a convolutional neural network (CNN) layer was also developed to assess the model's performance. Fusion architectures are a growing subset of deep learning architectures that uses the principle of combining data from multiple modalities (S. C. Huang, Pareek, Seyyedi, et al., 2020) and are proven to improve the result, causing expanded features (Ali et al., 2020; Khanna & Rana, 2020; Shi et al., 2022; H. Wang et al., 2020).

The contribution of this work is as follows: first, using feature extraction and Transformer based approach to improve and find the possible features that have a good correlation to represent information to the related epitope. The selected feature extractions were amino acid composition (AAC), dipeptide composition (DIP), Spectrum Descriptor, and a combination between AAC, Spectrum, and DIP called AADIP. The second, fusion of the Transformer with CNN to improve the classification performance is expected to allow the model to benefit from both methods, capture long-range dependencies and increase performance (Ali et al., 2020; S. C. Huang, Pareek, Seyyedi, et al., 2020).

Hence, it should be able to solve issues of low performance that previous studies encounter. We used multiple datasets from IEDB, VDJdb, and McPAS-TCR and used the following epitope: GILG, GLCT, and NLVP to evaluate our proposed model. In addition, this study will use the same AUC score metric as previous studies to evaluate the models for a fair comparison and show their robustness.