改进　采用基于 BERT 的深度学习方法改进癌症标志分类 h

Sultan ZAVRAK[1*]和 Seyhmus YILMAZ[1]

Sultan ZAVRAK

[1]杜兹杰大学计算机工程系，土耳其杜兹杰

sultanzavrak@duzce.edu.tr***ORCID***标识符： 0000-0001-6950-8927

Seyhmus YILMAZ[1]杜兹杰大学计算机工程系，土耳其杜兹杰

seyhmusyilmaz@duzce.edu.tr

ORCID标识符：0000-0001-9987-2797

***通讯作者：**

Sultan ZAVRAK

计算机工程系，工程学院，杜兹杰大学，81620，杜兹杰/土耳其 电话： +90 (380) 542 1036 电子邮箱：

sultanzavrak@duzce.edu.tr

改进　基于 BERT 深度学习方法的癌症标志分类　h

摘要

本文提出了一种能够准确分类癌症标志的新方法，这项任务对于癌症研究至关重要。我们提出的方法采用了基于变换器的双向编码器表示（BERT）架构，该架构已在各种下游应用中展现出卓越性能。通过应用迁移学习，我们在一个与癌症相关的小型生物医学文本文件语料库上对预训练BERT模型进行了微调。实验研究结果表明，我们的方法达到了94.45%的显著准确率，超过了几乎所有先前研究结果，并较文献报道的结果至少提高了8.04%。这些发现凸显了我们提出的模型在准确分类和理解癌症研究文本文件方面的有效性，从而为该领域作出了重要贡献。鉴于癌症仍是全球十大主要死因之一，我们的方法有望大力推动癌症研究并改善患者结局。

关键词：BERT、癌症标志分类、迁移学习、深度学习、自然语言处理

1. 引言

癌症是当今世界许多地区人们面临的最难治疾病之一，其中包括表观遗传和遗传突变（Jiang等人，2020年）。迄今为止，全球已有数百万人死于这种疾病（ Organization，2008年）。癌症研究历史悠久，从过去延续至今，并始终吸引着生物医学研究人员的关注。

在生物医学文献中，癌症常以其标志特征来描述。这些标志特征是一系列相互关联的生物学特征和行为，会促进癌症在体内生长（Baker、Korhonen等，2016年）。癌症标志是用于区分恶性细胞与健康细胞的特征（Yan & Wong，2017）。科学家可以借此更好地理解这些细胞活动如何导致癌症。癌症标志概括了用于界定疾病生长过程的细胞程序。在开创性论文（Hanahan & Weinberg，2000年）中，研究者首次提出六项癌症标志；随后，又在另一篇论文（Hanahan & Weinberg，2011年）中增加了四项，形成了目前公认的十项标志。迄今为止，研究者已探讨了对癌症研究及其疗法发现具有重大影响的十项癌症标志：（i）维持增殖信号（PS）、（ii）逃避免疫摧毁（ID）、（iii）逃避生长抑制因子（GS）、（iv）激活侵袭和转移（IM）、（v）促肿瘤炎症（TPI）、（vi）诱导血管生成（A）、（vii）使细胞获得复制永生（RI）、（vii）抵抗细胞死亡（CD）、（viii）失调细胞能量代谢（CE）、（ix）基因组不稳定性和突变（GI）（ Baker、Silins等，2016年）。

识别癌症标志对于癌症研究及其治疗探索至关重要。为了推动癌症研究的发展，能够通过自动将特定文章或文本归类到相应的癌症标志，预测其中是否包含特定的相关信息，具有极其重要的意义。

自然语言处理（NLP）的发展为解决此类问题提供了方案。文本理解与分类是从生物医学文本数据中识别癌症标志的一种方法，其中包括运用自然语言处理技术识别并分类生物医学文本信息。近年来，Transformer模型在特征提取方面的应用（如BERT）在许多文本任务中的表现显著优于以往的自然语言处理框架，例如情感分析（Xu等，2019年）、垃圾信息检测 （Yilmaz和Zavrak，2022年）等（Minaee等，2021年）（Zhou等，2022年）。这为改进稳健的分类器带来了巨大机遇，使其能够依据文本所涉及的标志特征进行准确分类。在此背景下，我们开发了一种基于先进自然语言处理方法（包括BERT）的检测模型，用于解决在生物医学摘要中识别癌症标志的文本分类问题。因此，本研究的主要目标之一是确定，在训练文本数据不足的情况下，基于BERT的架构能否提升癌症标志检测的性能。

近年来，预训练语言模型（如 BERT 和 ELMo）的应用在一系列文本分类任务中取得了显著进展，包括情感分析、情绪分类和主题分类（Devlin等，2018年）（Howard & Ruder, 2018）（Peters et al., 2018）（Radford et al., n.d.）。这一方法的核心理念是，基于大量语料构建的神经网络语言模型能够提供一种融合了不同分析层面信息的表示，例如语义、句法、形态和词汇层面（Barlas & Stamatatos, 2021年）。因此，这些模型可以在特定领域的语料上进行微调。从本质上说，这是一种迁移学习，因为将一个领域中获得的知识加以调整并应用于另一个领域的任务。

本文的一个重要目标是实施前沿的自然语言处理技术，以帮助未来的研究人员克服相关问题，并成功理解用于快速且稳健地分类癌症标志的自动化解决方案。更具体地说，本文研究使用深度学习（DL）进行生物医学文本分类，而不是采用手动特征工程。本研究使用Baker、Korhonen等人（2016年）的问题设置和数据集，但不同于卷积神经网络（CNN）、长短期记忆（LSTM）等方法，本文通过采用迁移学习，重点关注预训练BERT模型（Devlin等，2018年）。基于BERT的方法和“深度”网络广泛用于通用自然语言处理，但将其应用于生物医学文本的研究仍然很少。总体而言，尽管深度学习和基于BERT的技术，尤其是在通用领域的自然语言处理中，正逐渐得到普遍应用，但在生物医学文本中使用此类技术的研究相对较少。事实上，据我们所知，现有文献中尚无用于标志特征的基于BERT的模型。因此，本研究将首次使用基于BERT的模型来检测癌症标志。本文最重要的贡献总结如下：

- 基于BERT架构，使用一个包含1,800多篇标注了十种癌症标志的生物医学摘要语料库，构建了一个生物医学文本分类模型。• 对微调后的BERT模型的分类性能进行了评估，并与先前的文本分类方法进行了比较。• 实验结果表明，所提出的模型表现出显著改进，相较于先前的文本分类方法至少提升了8.04%，准确率达到94.45%。

- 这些结果证实了最优架构在有效解决摘要文本中癌症标志分类这一挑战方面的有效性。

本文组织如下。第2节介绍生物医学文本分类领域的相关工作。第3节阐述用于癌症标志分类的所提方法。第4节介绍公开可访问的标志特征数据集及实验设置，并讨论各模型的结果。最后一节给出总结性结论。

2. 相关工作

本节将概述生物医学文本分类领域开展的研究。具体而言，我们将首先讨论传统的机器学习方法，这些方法涉及手动创建特征；随后介绍采用特征学习而非特征工程的深度神经网络架构相关研究。

2.1. 传统方法

随着癌症相关生物医学非结构化和结构化数据规模的急剧增长，各种专注于从科学文献中发现信息的生物医学文本挖掘方法应运而生。研究人员已经发表了许多机器学习方法，包括朴素贝叶斯（Wang等人，2007年）、梯度提升（Jiang等人，2019年）、最大熵建模（Raychaudhuri等人，2002年）、（Shatkay等人，2008年）、支持向量机（SVMs）（Cohen，2006年；Garla等人，2013年）等，以推动癌症研究的发展。另一方面，正如下面将详细说明的，迄今为止，关于癌症标志文本分类的出版物数量仍然十分有限。

Baker等（Baker、Silins等，2016年）介绍了一个数据集，其中包含来自生物医学出版物的1,800多篇摘要，这些摘要按照癌症的十种标志特征进行了标注。此外，他们还实现了一种基于机器学习的癌症标志分类技术。该技术采用传统的自然语言架构，从中提取一组特定特征，用于训练支持向量机（SVM）分类器。特征形式多种多样，例如语法关系、动词类别、命名实体、词形还原词袋、医学主题词、化学物质列表以及名词二元组。该方法通过检测癌症获得了不错的结果，平均 F 分数为 77%。另一方面，其主要缺点是采用复杂的自然语言架构会带来高昂成本和计算需求。

Yan等人关注维数灾难，并使用1,580篇PubMed摘要比较了多种癌症标志标注技术（Yan & Wong，2017）。他们实现了一种名为UDT-RF的新方法，该方法利用本体特征。该方法使用医学主题词（MeSH）本体图来扩展特征空间。此外，他们使用多种性能指标对先进技术进行比较和评估，从而揭示这些技术在完整癌症标志集合上的整体性能范围。为展示所提方法如何揭示癌症研究的新见解，他们开展了若干案例研究。

然而，手动生成一组特征耗时耗力，也无法保证取得更好的结果（Jiang等人，2020年）。尽管采用了特征筛选方法，但随着特征集规模扩大，筛选出有益且重要的特征会变得更加困难。

2.2. 深度学习方法

如今，基于深度神经网络（DNN）的方法已在多种自然语言处理（NLP）应用中取得了先进成果，例如机器翻译、图像分类、语义相似性、释义识别、语言建模、问答、文档摘要和观点挖掘（Jiang等人，2020年）。

当深度神经网络在其他领域取得巨大成就时，研究人员开始在生物医学领域使用深度学习，这些应用包括关系抽取 (Gao 等，2019年；Hanahan & Weinberg，2000年、2011年)、概念抽取与编码（Garg 等，2019年）、表型分析 (Hochreiter & Schmidhuber，1997年) 以及命名实体识别（Bai 等，2016；Lazebnik，2010年）。关于癌症标志文本分类，文献中使用深度神经网络的研究仍然有限。

在（Baker、Korhonen 等，2016）中，作者重点研究了通过深度神经网络方法对生物医学文档进行分类，该方法强调特征学习，而不是手工设计的特征工程。他们使用了（Baker、Silins 等，2016年）作者引入的数据集，并且也采用了（Baker、Silins 等，2016年）中的相同问题设置。不过，他们更重视卷积神经网络，而不是支持向量机。实验结果表明，某些卷积神经网络初始化方法、训练过程和超参数能够使网络取得优于先前提出的支持向量机模型的结果；后者使用了手工构建的工程特征，其中包含命名实体识别和句法分析的输出。

为了利用标签共现关系（例如上位关系），(Baker & Korhonen, 2017) 的作者提出了一种用于层次化多标签文档分类的新颖技术。该技术在 Kim（Kim, 2014）的网络所采用的基础卷积神经网络框架中初始化最后一个隐藏层，而该框架通常应用于文本分类任务。他们利用生物医学领域中的两个多标签分类问题对该方法进行了探索，同时采用文档级分类和句子级分类。在实验设置中，该技术取得了令人鼓舞的性能表现。

Jiang等人（Jiang等人，2020年）提出并实现了一种名为深度上下文化注意力双向长短期记忆网络（DECAB-LSTM）的新模型，用于癌症标志文本分类。该方法采用上下文注意力机制，以学习捕获句子中最重要的部分。此外，还探讨了优质词嵌入对癌症标志文本分类的影响。在监督学习环境下，研究人员在一个基准数据集上开展了一些实验。他们利用癌症的十种标志，通过准确率、宏平均F1分数和曲线下面积分数等多种评估指标来测试其方法。在实验过程中，该模型在癌症标志文本分类方面取得了优于基线架构的先进结果。

BERT（基于变换器的双向编码器表示）是一种预训练语言模型，在多种文本分类任务中实现了最先进性能，彻底改变了自然语言处理（NLP）任务。因此，采用 BERT架构，有望在癌症标志分类方面取得比先前研究中提及的传统方法和深度学习方法更好的结果。此外，BERT 模型能够捕捉生物医学文本数据中不同概念之间复杂而细微的关系，因此可以更准确、更稳健地对癌症标志进行分类。

3. 所提出的模型架构

本研究采用 BERT 模型。该模型是一种最先进的双向 Transformer 架构，已在一系列自然语言处理（NLP）应用中展现出令人瞩目的性能（Qasim 等，2022年）。

3.1. BERT

BERT 网络架构最初由 Devlin 等人提出（Devlin等，2018年），是一种多层双向 Transformer 编码器，基于 Vaswani 等人的原始模型（Vaswani等人，未注明日期）。BERT（Devlin等，2018年）是一种简洁而强大的模型，在众多机器学习应用中取得了显著的改进。BERT 的主要目标是从无标签数据集中学习双向表示，并利用所有层中的左右文脉协同作用（Qasim 等，2022年）。

预训练和微调是 BERT（Devlin等，2018年）采用的两个不同步骤。在预训练阶段，使用无标签数据在各种预训练任务上训练网络。首先，使用预训练参数初始化 BERT 网络；随后，利用取自下游自然语言处理任务的带标签数据集，对全部参数进行微调。尽管这些网络都以相同的预训练参数为起点，但每个下游问题都有各自不同的微调网络。BERT 的一项显著特点是，它能在不同的下游应用中采用统一的模型（Devlin等，2018年）。最后的下游模型与预训练模型之间的差异极小。

在本研究中，我们使用 BERT-base 模型，并将模型中的自注意力头数量表示为 A、隐藏维度表示为 H、层数（即 Transformer 块数量）表示为 L。BERT 模型在标志分类任务中的框架如图 1 所示。

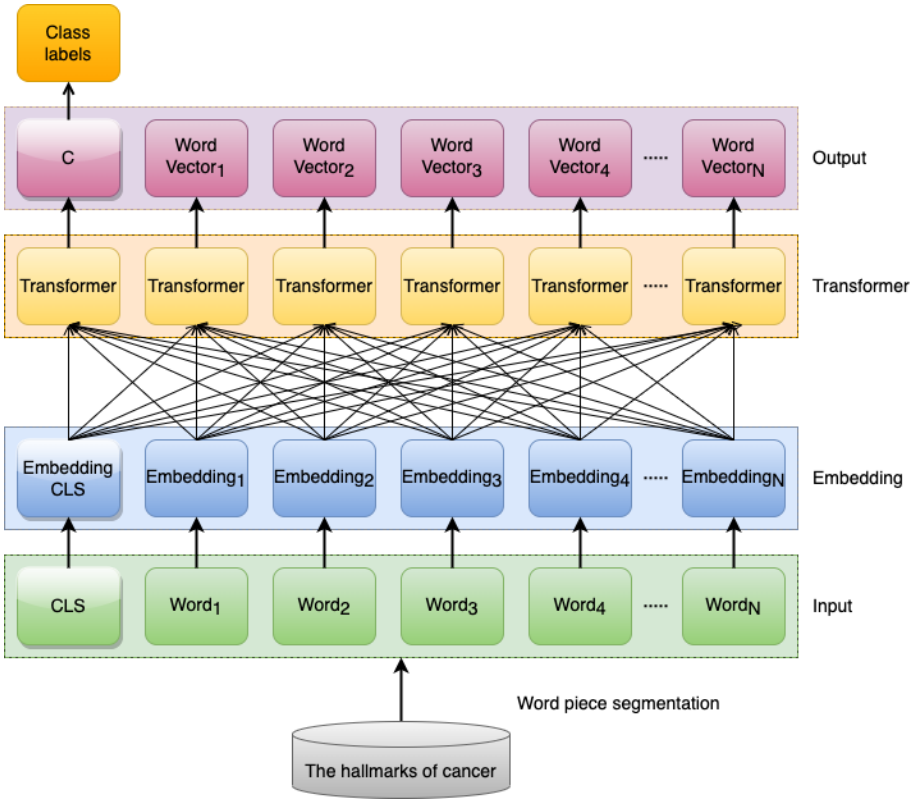

图 1. 标志分类任务中 BERT 模型的框架（Xiaofeng 等人，2021年）

最初，Transformer 被引入作为特征提取器，以提升翻译任务的性能。后来，研究人员证明，在引入自注意力机制和位置编码后，Transformer 也能高效地应用于各种人工智能领域。Transformer 使用编码器-解码器结构。解码器和编码器在架构上相同，且都包含六个相似的层。解码器中设有一个额外的注意力子层，而编码器中没有。每层包含两个子层：一个全连接前馈神经网络层和一个多头自注意力层（Xiaofeng 等人，2021年）。编码器的架构如图 2 所示。

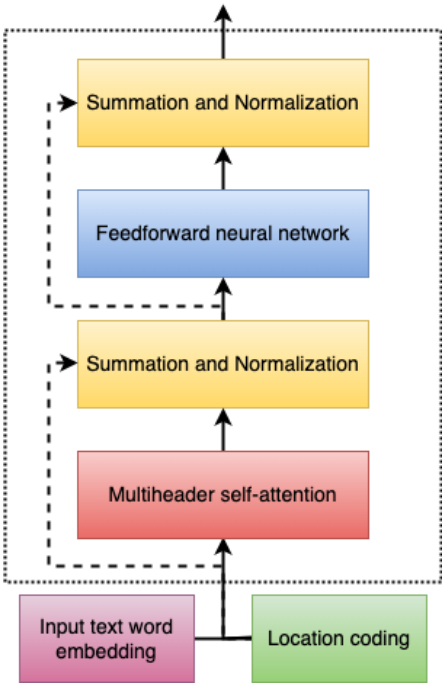

图 2. 编码器的架构（Xiaofeng 等人，2021年）

多头自注意力层是若干自注意力架构的组合，也是 Transformer 中最重要的组成部分之一。编码器和解码器都包含多头自注意力层。最初提出注意力机制时，解码器负责处理训练网络时需要翻译的文本文件部分。为了预测输出，系统会根据注意力机制计算不同输入的权重。自注意力机制计算目标文本与原始文本之间的注意力。另一方面，注意力机制也计算目标文本与原始文本之间的注意力。这提升了模型识别相距较远的句子成分之间特征的能力。这样一来，网络便能捕捉文本中相距较远的元素之间的特征。注意力机制的矩阵计算公式见公式 1（Xiaofeng 等人，2021年）。

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad 1$$

根据上述公式，V 表示值矩阵，K 表示键矩阵，Q 表示查询矩阵。这些值是文本嵌入矩阵与各自权重矩阵相乘所得的结果。查询向量的维度用 d_k 表示。基于此，与单头自注意力机制相比，多头自注意力机制能够获得更优的性能。为了增强网络对不同位置词语的关注能力，模型使用了多个键权重矩阵、查询权重矩阵和值权重矩阵。其计算公式如公式 2 所示（Vaswani等人，n.d.；Xiaofeng 等人，2021年）：

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^O \quad 2$$

根据上述公式，W^O 表示联合训练的权重矩阵，而通过第i个注意力头获得的权重矩阵投影矩阵为 head_i =

$Attention(QW_i^Q, KW_i^K, VW_i^V)$，$W_i^Q$、$W_i^K$ 和W_i^V。为了指定当前词的编码，自注意力机制允许网络在自然语言处理过程中获取文本中的其他词作为参考因素，从而对一个词进行编码（Vaswani等人，n.d.）。仅依靠自注意力机制无法表示输入的绝对/相对位置顺序信息，尽管在处理文本文件时，位置信息对每个词元都非常重要（Xiaofeng 等人，2021年）。因此，为了解决这一问题，在 Transformer 中将一个向量添加到每个输入词元的嵌入中，以通过位置编码等方式指定其位置信息。需要计算的公式如公式3和4所示。

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{model}}}\right) \quad 3$$

$$PE_{(\text{pos},2i+1)} = \cos\left(pos/10000^{2i/d_{\text{model}}}\right) \quad 4$$

在公式 3 和 4 中，i 表示维度，pos 表示文本中词元的位置信息。若将$pos = 0$、1、2、...、$L-1$ 和L 视为句子的长度，词模型中的向量长度记为$dmodel$，位置编码向量中第i个维度由i 表示，因此$i = 0$、1、2、...、$(dmodel/2-1)$。因此，BERT架构中的 Transformer 利用位置嵌入来表示文本文件中的位置信息，并改进了原始的位置编码（Xiaofeng 等人，2021年）。此外，为了在训练时尽可能多地捕获文本中的上下文数据，BERT采用了双向架构。

使用 BERT 进行词嵌入时，会生成 BERT 的输出向量。BERT-base架构包含 12 个 Transformer 编码器块（Park等，2022年）。Transformer 编码器块包含前馈网络、归一化和自注意力组件。经过字节对编码（BPE）分词器处理的一系列词语作为输入X。输入X中词语之间的相关性由Z 表示。通过加权求和和缩放点积注意力机制计算Z。自注意力机制由公式 5 至公式 11 表示（Park等，2022年）。

$$X = [x_1, x_2, \cdots, x_n] \quad 5$$

$$Q = W_q \times X = [q_1, q_2, \cdots, q_n] \quad 6$$

$$K = W_k \times X = [k_1, k_2, \cdots, k_n] \quad 7$$

$$V = W_v \times X = [v_1, v_2, \cdots, v_n] \quad 8$$

$$Z = softma\,x\left(\frac{QK^t}{\sqrt{d_k}}\right)V = [z_1, z_2, \cdots, z_n] \quad 9$$

$$\dot{Z} = Normalize\ (X + Z) \quad 10$$

$$\ddot{Z} = Normalize\left(X + FFN(\dot{Z})\right) \quad 11$$

如式（12）和式（13）所示，通过归一化和前馈步骤计算出最后一个向量 $\ddot{Z}$。上述过程重复几次后，最终层的 $\ddot{Z}$ 被用作词嵌入。每个嵌入向量都取决于包含该词的文本。因此，这类向量称为上下文嵌入。

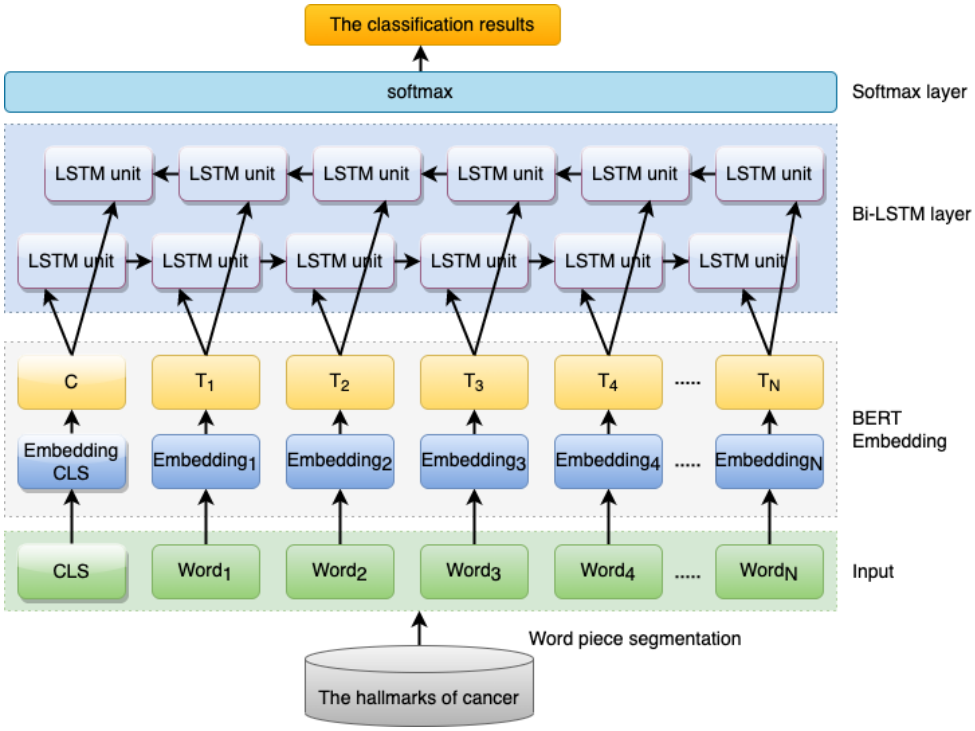

图3. 模型结构示意图

预训练和微调是 BERT 使用的两个不同步骤（Devlin等，2018年）。在预训练阶段，使用无标签数据，在各种预训练任务上训练网络。首先，使用预训练参数初始化 BERT 网络；随后，利用来自下游自然语言处理任务的带标签数据集，对全部参数进行微调。尽管它们都使用相同的预训练参数进行初始化，但在每个下游问题中都会得到不同的微调网络。

在 BERT 中，微调是一种监督学习（Park等，2022年），用于文本分类等领域。通常，当 BERT 应用于情感分析和问答等任务时，微调会使用 [CLS], ，这是一个特殊符号，会作为输入放在每个输入样本的开头（Devlin等，2018年）。[CLS] 表示用于计算标签的分类似然。池化操作应用于 CLS 符号的最后一个隐藏状态。通过分类层应用 SoftMax函数，计算出似然最大的标签。随后对分类层和预训练 BERT 参数进行微调，以估计正确标签的对数似然。

4. 实验

4.1. 数据集描述

本研究采用了一个包含 1852 篇生物医学出版物摘要的语料库，这些摘要已根据文献中记载的癌症标志进行了标注（Baker、Silins等，2016年）。在本研究中，所提出的模型使用该语料库进行训练和评估。数据集中的每篇摘要都可能与所研究的十种癌症标志中的零种或多种相关联。因此，本研究的目标涉及多标签分类，这种方法将在下文中作更详细的阐述（Zhou等，2022年）。

维持增殖信号：正常细胞需要信号分子才能生长和分裂。与正常细胞不同，恶性细胞无需这些信号也能生长。

逃避免疫生长抑制：细胞中存在一些阶段，会阻止细胞生长和分裂。癌细胞中的这些阶段发生了改变，因此无法有效阻止细胞分裂。

抵抗细胞死亡：细胞受损时会自动死亡，这一过程由称为细胞凋亡的机制控制。恶性细胞可以克服这种机制。

使细胞获得复制永生：正常细胞分裂到一定次数后就会死亡。但恶性细胞能够无限生长和分裂，这种现象称为永生。

诱导血管生成：恶性细胞可以启动一个称为血管生成的过程，从而形成新的血管，为其提供所需的氧气和其他营养物质。

激活侵袭与转移：恶性细胞可以离开最初所在的位置，侵占邻近组织，并扩散到身体较远的部位。

基因组不稳定性与突变：恶性细胞中通常会出现严重的染色体异常，而且随着疾病进展，异常情况会进一步恶化。

Avo免疫清除：免疫系统无法识别恶性细胞。

失调细胞能量代谢：许多恶性细胞会利用异常的代谢通路产生能量，例如，即使氧气充足、能够正常吸入氧气时，也会进行葡萄糖发酵。

促肿瘤炎症：炎症通过影响肿瘤附近的微环境，促进恶性细胞的增殖、转移和存活。

表 1. 数据集的标注统计

	训练集 (70.36%)		验证集 (9.88%)		测试集 (19.76%)		总计	
标志特征	**pos**	**neg**	**pos**	**neg**	**pos**	**neg**	**pos**	**neg**
维持增殖信号	328	975	43	140	91	275	462	1390
逃避免疫生长抑制	172	1131	22	161	46	320	240	1612
抵抗细胞死亡	303	1000	42	141	84	282	429	1423
使细胞获得复制永生	81	1222	11	172	23	343	115	1737
诱导血管生成	99	1204	13	170	31	335	143	1709
激活侵袭和转移	208	1095	29	154	54	312	291	1561
基因组不稳定性和突变	227	1076	38	145	68	298	333	1519
促肿瘤炎症	169	1134	24	159	47	319	240	1612
细胞能量代谢	74	1229	10	173	21	345	105	1747
逃避免疫摧毁	77	1226	10	173	21	345	108	1744

4.2. 实验设置

在监督学习领域，我们在一个基准数据集上开展了一系列严格的实验，其中每个文本序列都经过人工标注，以赋予其相应的标志特征。我们使用训练集划分来训练模型，并选择了最具潜力的模型

使用开发（验证）集选择了该参数。随后，使用测试集全面评估模型的效果和性能。

作为一个研究平台，本研究采用 Google Colab Pro，该平台基于 Tesla P100 GPU 提供专用的图形处理环境，用于开展实验。研究使用 ktrain（Maiya，2022年）这一面向 TensorFlow Keras 的轻量级封装库，训练和测试深度学习与机器学习模型。

针对癌症标志的分类，预训练的 BERT-base 最初以相对较少的参数加载，并在以 TensorFlow 为后端的 Keras 上运行。本研究采用了 BERT-base 预训练模型的有大小写和无大小写两种配置。基础模型由 12 个 Transformer 块层组成，隐藏维度为 768，自注意力头数量为 12，总计包含 1.1 亿个参数。词元的最大长度限制设为 512。批量大小设为 6，以便一次提供较小的训练集。在训练策略中采用 1 周期策略（Smith，2018年），用于确定紧密耦合的学习率、动量和正则化的最优值。我们启动 fit 函数进行 20 个训练轮次，同时将学习率保持在 1e-5。

4.3. 结果与讨论

鉴于样本分布存在显著差异，本文基于宏平均F1值、宏平均精确率、宏平均召回率和分类准确率等指标，并结合曲线下面积，对所提出的模型在十种癌症标志上的表现进行了评估（Song等，2022年）。这些模型的评估结果见表2、表3和表4。

表2. 微调后的 BERT-base-uncased 模型结果（%）

标志特征	准确率	宏平均精确率	宏平均召回率	宏平均F1值	曲线下面积
维持增殖信号	93.44	85.36	91.56	88.04	97.17
逃避免疫生长抑制	97.81	89.90	89.90	89.90	99.50
抵抗细胞死亡	99.18	93.75	99.57	96.45	99.79
使细胞获得复制永生	99.18	97.44	95.51	96.45	98.11
诱导血管生成	90.98	78.92	84.61	81.36	93.64
激活侵袭和转移	92.90	87.71	89.39	88.52	94.85
基因组不稳定性和突变	95.36	82.87	93.07	87.07	94.68
促肿瘤炎症	95.63	92.34	96.33	94.11	97.32
细胞能量代谢	81.97	76.06	79.18	77.30	88.62
逃避免疫摧毁	96.99	92.37	94.65	93.46	97.19
平均值	94.34	87.67	91.38	89.26	96.09

表2展示了微调后的BERT-base-uncased模型的结果，表明其在各项分类任务中的准确率介于81.97%至99.18%之间。同样，表3展示了微调后的BERT-base-cased模型的结果，表明其在各项分类任务中的准确率介于88.29%至99.82%之间。平均而言，BERT-base uncased模型的表现优于BERT-base cased模型，其召回率、F1值和曲线下面积分数更高。相反，在准确率和精确率方面，BERT-base cased模型的表现优于BERT-base uncased模型。

表 3. 所提出的 BERT-base-cased 模型微调后的结果（%）

标志特征	准确率	宏平均精确率	宏平均召回率	宏平均F1值	曲线下面积

维持增殖信号	93.44	85.36	91.56	88.04	96.97
逃避免疫生长抑制	98.09	91.92	90.04	90.96	98.87
抵抗细胞死亡	99.18	93.75	99.57	96.45	99.82
使细胞获得复制永生	99.18	95.69	97.53	96.59	98.44
诱导血管生成	90.98	79.57	79.02	79.29	89.36
激活侵袭和转移	92.62	87.96	87.52	87.74	95.96
基因组不稳定性和突变	96.17	85.88	92.06	88.65	96.02
促肿瘤炎症	94.81	91.77	94.12	92.87	97.01
细胞能量代谢	83.61	78.00	81.00	79.23	88.29
逃避免疫摧毁	96.45	91.76	92.52	92.14	96.05
平均值	94.45	88.17	90.49	89.20	95.68

表4对比了以准确率、精确率、召回率、F1值和曲线下面积为指标的既有研究中各种方法与微调后的BERT模型的表现。结果表明，除曲线下面积外，微调后的BERT模型在所有指标上均优于既有研究。微调后的BERT-base-uncased模型准确率达到94.34%，显著高于既有研究报告的最高准确率（DECAB-LSTM为86.3%（Jiang等人，2020年））。同样，微调后的BERT-base-cased模型准确率达到94.45%，也高于既有研究取得的最高准确率。

表4. 与先前研究的比较（%）

Ref.	**方法**	**准确率**	**宏平均-精确率**	**宏平均-召回率**	**宏观 F1**	**AUC**
Baker，Korhonen，和 Pyysalo（Baker, Korhonen，等人，2016年）	经过调优的卷积神经网络	-	-	-	81.0	97.6
Jiang等人（Jiang等人，2020年）	DECAB-LSTM	86.3	-	-	89.10	98.9
Prabhakar 和 Won (Prabhakar & Won，2021年)	混合双向门控循环单元与多头注意力	74.71	70.82	68.99	69.89	-
本研究	微调后的 BERT-base-uncased	94.34	87.67	91.38	89.26	96.09
	微调后的 BERT-base-cased	94.45	88.17	90.49	89.20	95.68

微调后的BERT模型表现更优，这可归因于它们在文本数据上经过了大量预训练，因此对语言及其细微差异有更深入的理解。此外，BERT模型采用双向Transformer，能够捕捉词语在两个方向上的关系，而卷积神经网络和长短期记忆网络等传统模型则无法做到这一点。此外，使用生物医学文本数据对这些模型进行微调，能够提升它们对该语境的理解，使其从文本数据中提取更多上下文信息和含义，这对于准确分类癌症标志至关重要。

5. 结论

总之，本文提出了一种使用 BERT（一种双向Transformer模型）对生物医学文本进行分类的新方法。该方法采用癌症标志

进行评估。这些标志特征在癌症研究中已变得极其重要。实验结果表明，所提出的方法准确率达到 94.45%，优于目前文献中几乎所有已发表的结果。

这项工作的贡献意义重大，因为尽管基于深度神经网络的方法在各种自然语言处理领域日益普及，但将这一策略应用于生物医学文本，尤其是癌症标志的研究仍然相对较少。所提出的模型填补了这一空白，并展示了将深度学习用于癌症研究的潜力。

尽管所提出的模型取得了较高的准确率，但这项工作仍存在一些局限性。首先，所提出的模型仅专注于癌症标志分类，仍需进一步研究其在其他类型生物医学文本上的表现。其次，本研究使用的数据集规模相对较小，可能无法代表所有癌症类型。

未来的工作可以通过采用规模更大、更多样化的数据集，并探索所提出模型在其他类型生物医学文本上的表现，来解决这些局限性。此外，还可以进一步研究结合不同深度学习技术以提高所提出模型准确率的潜力。

数据可用性

本研究使用的数据集取自此前发表的研究成果。

竞争性利益

作者声明，他们不存在任何利益冲突。

资金支持

不适用。

作者贡献

每位作者对本研究的概念化、方法论、分析和写作贡献相同。

参考文献

Bai, T., Gong, L., Wang, Y., Wang, Y., Kulikowski, C. A., & Huang, L. (2016). 一种用于探索生物医学知识网络中隐式概念相关性的方法。*BMC*生物信息学，*17*(9), 53– 66. https://doi.org/10.1186/S12859-016-1131-5/TABLES/3

Baker, S., & Korhonen, A. (2017). 用于分层多标签文本分类的神经网络初始化。*2017*年生物医学自然语言处理研讨会（*BioNLP*）——*SIGBioMed* 生物医学自然语言处理研讨会，第*16*届 *BioNLP* 研讨会论文集，307–315. https://doi.org/10.18653/V1/W17-2339

Baker, S., Korhonen, A., & Pyysalo, S. (2016). 使用卷积神经网络进行癌症标志文本分类. 1–9. https://cambridgeltl.github.io/cancer-hallmark-cnn/

Baker, S., Silins, I., Guo, Y., Ali, I., Högberg, J., Stenius, U., & Korhonen, A. (2016). 自动语义 c
根据癌症标志对科学文献进行分类。《生物信息学》, *32*(3), 432–440.
https://doi.org/10.1093/BIOINFORMATICS/BTV585 Barlas, G., & Stamatatos, E. (2021). 跨领域作者归属判定的迁移学习方法。《演化系统》, *12*(3), 625–643.
https://doi.org/10.1007/S12530-021-09377- 2/FIGURES/5 Cohen, A. M. (2006). 一种有效的通用自动化生物医学文档分类方法。美国医学信息学会年度研讨会论文集, *2006*年, 161. / pmc/articles/PMC1839342/ Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert：用于语言理解的深度双向变换器预训练。*arXiv*预印本 *ArXiv:1810.04805*. Gao, Z., Fu, G., Ouyang, C., Tsutsui, S., Liu, X., Yang, J., Gessner, C., Foote, B., Wild, D., Ding, Y., & Yu, Q. (2019). Edge2vec：利用边语义进行表示学习，以实现生物医学知识发现。*BMC*生物信息学, *20*(1), 1–15. https://doi.org/10.1186/S12859-019-2914- 2/FIGURES/6 Garg, S., Galstyan, A., Steeg, G. ver, & Cecchi, G. (2019). 用于生物医学关系抽取的近乎无监督哈希码表示。*2019*年*EMNLP-IJCNLP*——*2019*年自然语言处理经验方法会议暨第九届自然语言处理国际联合会议论文集, 4026–4036. https://doi.org/10.18653/V1/D19-1414 Garla, V., Taylor, C., & Brandt, C. (2013). 使用拉普拉斯支持向量机进行半监督临床文本分类：在癌症病例管理中的应用。《生物医学信息学杂志》, *46*(5), 869–875.
https://doi.org/10.1016/J.JBI.2013.06.014 Hanahan, D., & Weinberg, R. A. (2000). 癌症标志。《细胞》, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9 Hanahan, D., & Weinberg, R. A. (2011). 癌症标志：新一代。《细胞》, *144*(5), 646– 674.
https://doi.org/10.1016/J.CELL.2011.02.013 Hochreiter, S., & Schmidhuber, J. (1997). 长短期记忆。《神经计算》, *9*(8), 1735

–
1780. https://doi.org/10.1162/NECO.1997.9.8.1735 Howard, J., & Ruder, S.（2018年）。用于文本分类的通用语言模型微调。*ACL 2018* - 第*56*届计算语言学协会年会论文集

e
会议（长篇论文），*1*，328–339。https://doi.org/10.48550/arxiv.1801.06146 Jia ng, L.、Sun, X.、Mercaldo, F. 和 Santone, A.（2020）。DECAB-LSTM：用于癌症标志分类的深度上下文化注意力双向长短期记忆网络。基于知识的系统，*210*，106486。
https://doi.org/10.1016/J.KNOSYS.2020.106486 Jiang, L.、Zhang, B.、Ni, Q.、Sun, X. 和 Dong, P.（2019）。基于基尼不纯度的梯度提升法预测单核苷酸多态性序列。*IEEE Access*，*7*，12647–12657。https://doi.org/10.1109/ACCESS.2019.2893269 Kim, Y.（2014）。用于句子分类的卷积神经网络。载于*201*自然语言处理实证方法会议（*EMNLP*）论文集。计算语言学协会。https://doi.org/10.3115/v1/d14-1181

4

Lazebnik, Y. (2010)。癌症标志是什么？《自然综述：癌症》，*10*年(4)，232–233。https://doi.org/10.1038/NRC2827 Maiya, A. S. (2022)。ktrain：用于增强机器学习的低代码库。《机器学习研究杂志》，*23*，1–6。https://github.com/Tony607/Chinese_sentiment_analysis Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021)。基于深度学习的文本分类。《*ACM*计算综述》（*CSUR*），*54*(3)。https://doi.org/10.1145/3439726 Organization, W. H. (2008)。《*2008*年世界卫生统计》。世界卫生组织。https://apps.who.int/iris/handle/10665/43890 Park, Y., Ko, Y., & Seo, J. (2022)。在对话系统中使用话语注意力机制进行基于BERT的响应选择。《应用专家系统》，*209*，118277。https://doi.org/10.1016/J.ESWA.2022.118277 Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018)。深度上下文化词表示。*NAACL HLT 2018 - 2018*年北美计算语言学协会人类语言技术会议论文集，*1*，2227–2237。https://doi.org/10.18653/v1/n18-1202 Prabhakar, S. K., & Won, D. O. (2021)。使用带多头注意力的混合深度学习模型进行医学文本分类。《计算智能与神经科学》，*2021*年。https://doi.org/10.1155/2021/9425655 Qasim, R., Bangyal, W. H., Alqarni, M. A., & Ali Almazroi, A. (2022)。一种用于文本分类的基于微调BERT的迁移学习方法。《医疗保健工程杂志》，*2022*年。https://doi.org/10.1155/2022/3498123 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (无日期)。语言模型是无监督多任务学习器。检索于2022年8月15日，网址：https://github.com/codelucas/newspaper Raychaudhuri, S., Chang, J. T., Sutphin, P. D., & Altman, R. B. (2002)。通过对生物医学文献进行最大熵分析，将基因与基因本体代码关联起来。《基因组研究》，*12*(1)，203–214。https://doi.org/10.1101/GR.199701 Shatkay, H., Pan, F., Rzhetsky, A., & Wilbur, W. J. (2008)。生物医学文本的多维分类：面向多元用户自动化、实用地提供高效用文本。《生物信息学》（英格兰牛津），*24*(18)，2086–2093。https://doi.org/10.1093/BIOINFORMATICS/BTN381 Smith, L. N. (2018)。神经网络超参数的系统化方法：第*1*部分——学习率、批量大小、动量和权重衰减。https://doi.org/10.48550/arxiv.1803.09820 Song, R., Giunchiglia, F., Shen, Q., Li, N., & Xu, H. (2022)。利用在线交互网络改进辱骂性语言检测。《信息处理与管理》，*59*(5)，103009。https://doi.org/10.1016/J.IPM.2022.103009 Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (无日期)。《注意力就是你所需要的一切》。https://doi.org/10.5555/3295222.3295349

Wang, P.、Morgan, A. A.、Zhang, Q.、Sette, A. 与 Peters, B.（2007）。自动化文档分类 n 用于免疫表位数据库。*BMC*生物信息学，*8*(1)，1–10。 https://doi.org/10.1186/1471-2105-8-269/FIGURES/4 Xiaofeng, W.、Jinghua, Z.、Chenxi, J. 和 Yiying, J.（2021年）。基于 BERT 的期货预测文本情感分类研究。《计算》，1–18。https://doi.org/10.1007/S00607-021- 00989-9/TABLES/11 Xu, H.、Liu, B.、Shu, L. 和 Yu, P. S.（2019年）。面向评论阅读理解和基于方面的情感分析的 BERT 后训练。*NAACL HLT 2019* - 北美计算语言学协会人类语言技术 *2019* 年会议论文集，*1*，2324–2335。https://doi.org/10.48550/arxiv.1904.02232 Yan, S. 和 Wong, K. C.（2017年）。通过富集本体阐明高维癌症标志注释。《生物医学信息学杂志》，*73*，84–94。 https://doi.org/10.1016/J.JBI.2017.07.011 Yilmaz, S. 和 Zavrak, S.（2022年）。使用上下文化词表示进行巨魔推文检测。https://doi.org/10.48550/arxiv.2207.08230 Zhou, B.、Zou, L.、Mostafavi, A.、Lin, B.、Yang, M.、Gharaibeh, N.、Cai, H.、Abedin, J. 和 Mandal, D.（2022年）。VictimFinder：利用 BERT 从社交媒体中收集灾害应急响应中的救援请求。《计算机、环境与城市系统》，*95*，101824。 https://doi.org/10.1016/J.COMPENVURBSYS.2022.101824