## 摘要

## Abstract

## 目录

## 第一章 绪论

### 1.1 研究背景

#### 1.1.1 商飞C919民航客机

在二十一世纪的今天，我国的经济发展势头迅猛，已经进入了经济发展的新时代，不再是改革开放初期百废待兴的情形，而是进入了改革的深水区。在这种形势下，为了进一步增强综合国力，提高国际竞争力，我国亟需完成的任务之一便是进行产业升级。当前我国的经济主体大部分仍是轻工业、加工业和小商品制造业等，这些产业利润低层次低，且缺乏国际影响力。相对之下，高科技产业、高端制造业等产业利润巨大且不可替代性强，具有很高的战略地位，掌握这些核心科技与技术能够使我国拥有更多的国际话语权，正是当今我国需要大力扶持的产业。航空制造业正是这种高层次产业之一。

研制和发展大型客机是建设创新型国家，提高我国自主创新能力和增强国家核心竞争力的重大战略举措。在经济全球化的背景下，世界各国的经济贸易往来非常频繁，这对远距离跨国、跨洋航空运输行业是十分有利的，即使在国内，远途出行的情况下，消费者也往往倾向于选择乘坐民航飞机出行，因此，民航飞机制造公司的订单量逐年递增。此外，大型民航飞机制造产业非常依赖国家的工业水平与高新科学技术，一架自主研发的大型民航客机不仅仅是一个国家轻重工业生产能力与高科技研发水平凝聚的结晶，是一个国家综合国力的体现，更是可以振兴国家的航空制造产业，为国家带来巨大的利润、外汇，增强综合国力，提高国家人民的生活水平。但是，由于大型民航飞机制造的技术壁垒存在，不管是国内市场还是国外市场，大型民航飞机的主要生产商均是空客公司与波音公司。如图所示（中国内地民用客机制造商比例分布图），空客与波音这两家外国公司垄断了国内的大型民航客机几乎95%的生产订单，所以为了不再每年耗费大量财力购买外国民航客机，也为了在大型民航客机制造领域不再受制于人，我国自主研发生产的一款大型民航客机C919就在万众瞩目中应运而生了。



C919大型民航客机由中国商用飞机有限责任公司（以下简称商飞）设计研发并生产测试。C919为国产中短程干线客机，座位数量为168座（基本型布局），拥有长达4075公里的标准航程，航程最大可达到5555公里，飞行时间可超过9万个小时，与空客A320、波音737属于同一级别的飞机。自主研发C919大型民航客机项目可以带来很大的收益，从政治的角度来说，可以提升我国的综合国力，提升我国在国际上的话语权等等；从经济的角度来说，可以节省采购进口外国公司民航客机的资金，节省巨量的外汇，还可以带动发展国内的民航客机产业链，创造大量的就业机会，拉动地区经济发展等等；从技术的角度来说，自主研发成功C919大型民航客机，实现从零到一的突破，这给国内航空行业内的科研人员、工程师、学者带来了许多宝贵的实践机会，可以启动大量基于民航客机各个方面的研究项目，并实际投入测试评估，以提升C919大型民航客机的综合性能。

#### 1.1.2 大数据机器学习

当今社会每天每时每刻都在产生巨量的信息与数据，这些巨量数据的量级一般是TB、PB级，在某些场景下甚至会出现EB级的数据量。至于这些巨量数据的来源则是分布在各行各业，它们以各种各样的格式存储在硬盘中，这些巨量数据就是大数据。大数据是互联网行业里的石油，在各行各业内部有着重要的战略地位，公司内部的大数据一般都会受到保护严禁外泄。这是因为在这些巨量数据里隐藏着的是同样巨量且宝贵的信息，上到天体运行规律，下到消费者消费习惯，大数据里蕴含的信息量无法用只言片语概括。但是在几十年前，受限于计算机的算力水平与数字通信传输能力，以及缺乏高效的分析算法，人们没有能力与条件处理分析数据量多至TB、PB级的大数据，更不用说挖掘出这些巨量数据中隐藏的宝贵信息。那个时期研究人员往往需要花费几周甚至月余的时间，使用经典数学统计方法分析这些巨量数据。这种情形直到通信与计算硬件水平趋于成熟的今天才得以改善，有了硬件的算力与传输支持，同时各种高效的机器学习算法如雨后春笋般涌现，大数据这才渐渐崭露头角，为人们的生产与生活增添极大的助力。

正如上文所说，只有能够处理大数据的硬件还不够，还需要合适且高效的算法支持才可以，近年来火热的机器学习领域中正存在着许多适合处理大数据的分析算法。机器学习是一门多领域交叉的学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。机器学习是研究如何训练计算机使其能够模拟或者重现人类的学习行为，从而获取新的知识或技能，进而能够脱离人类达到能够自我不断优化所学知识的程度。大多数机器学习算法是基于统计和经验来训练计算机如何学习和分析，如果用于训练的数据量足够多，理论上可以得到一个在该数据领域具有很强学习与分析能力的模型，而且将分析巨量数据的任务交给算力强大的计算机来完成是十分适合的，可以收获很高的效率。因此，许多机器学习算法与大数据是相辅相成的一种关系。

C919大型民航客机的机载交流发电机电源在飞行过程中会产生TB乃至PB级的数据量，从这些电源数据中我们可以获得许多有价值的信息，例如飞机交流电源的工作状态就可以在这些电源信号数据中体现，通过有效的算法去分析这些巨量的电源信号数据，我们可以建立一种对机载交流发电机电源可能产生的异常的预警机制，从而对C919大型民航客机的飞行安全做出保障。本次研究的课题正是依托于商飞C919大型民航客机的项目，聚焦于C919大型民航客机机载发电机电源质量参数的研究，采用大数据机器学习的方法来对机载发电机电源质量参数进行异常检测与实时分析。

### 1.2 研究内容与意义

#### 1.2.1 课题研究目标

本次课题的研究目标是通过大数据机器学习算法手段，深入挖掘大量电源信号数据中的相互关联性与历史时间的关联性，设计并实现一种能够高速实时处理机载发电机电源信号数据的异常检测方法，实现高效、实时、可靠地预测分析电源质量参数，达到预警潜在故障、确保大型民航客机机载电气设备稳定与安全的目的。同时也能够解决大型民航客机飞行过程中大量电源信号数据处理时间长、分析手段少、异常状态定位慢等关键问题，为大型民航客机飞行过程提供有效的数据分析与异常定位工具。

#### 1.2.2 课题研究内容与意义

本次课题的研究内容是大型民航客机机载交流发电机电源品质参数的异常检测，电源品质参数包括电源的电压偏差、电流偏差、电压波动和闪变（浪涌电流）、基波（频率偏差）、谐波含量等若干方面。在这些电源品质参数中，电压和电流等时域中的参数利用传统检测方法，无论是阈值检测法还是直接观察实时电源信号波形图的方法，均能够较为容易地检测出参数的波动与异常状态。然而电源频域中的品质参数，如基波频率值、谐波的含量等，使用传统的阈值检测法或者直接观察频谱图像的方法都存在一些问题。比如阈值检测法的检测结果可能会存在较大的延时，这样检测结果就会没有参考价值，这部分在本文第三章有详细的阐述；直接观察频谱图像的方法更是效率极低且误报率较高。

因此，本次研究将对电源品质参数研究的重心放在了电源信号频域品质参数的异常检测方面，设计并实现了一种高效且实时的电源信号频域数据异常检测与预测分析方法和模型。电源信号频域部分的品质参数主要包括基波频率偏差与谐波含量两部分，本次研究的内容正是对这两种参数进行实时预测分析与异常检测。

机载交流发电机是飞机电源系统中最重要的部件，它的工作状态直接影响到整个飞机机载设备的运行状况和飞机自身的飞行安全。因此，研究飞机交流发电机的异常检测技术具有十分重要的意义。本次研究的主要内容是大型民航客机机载交流发电机电源信号频域品质参数的异常检测，频域品质参数主要是基波频率偏差与谐波含量，接下来简单介绍这两种品质参数的意义与重要性。

基波频率是机载发电机电源的主频率，是机载电气设备用电的主频率，保证其值稳定的重要性不言而喻。飞机交流发电机电源的谐波属于噪声的一种，谐波含量应当维持在一个很小范围内，单次谐波含量应当小于基波的4%。尽管当前飞机电源普遍采用三相交流发电机供电，而且对于平衡的三相发电机而言，偶次谐波可以被消除，但是仍然存在若干含量较高的奇次谐波。如果基波频率值发生较大偏差，或者谐波含量超出标准范围，就会影响到飞机交流发电机的电源质量，严重时会导致电源品质严重下降，进而导致其他机载用电设备的工作异常，后果十分严重。因此，研究一种面向电源信号频域品质参数的高效实时异常检测方法是十分必要的。

本次研究提出了一种面向频域频率参数的实时预测与异常检测方法，同时设计实现了一种基于时频转换与LSTM神经网络的频域频率参数预测分析模型。经过本次研究对该模型进行理论分析与实验验证，最终的实验结果表明了该模型相较于当前其他方法与模型具有较大的性能提升。本次实验结果表明，该模型在频域方面对频率参数的预测有着较高的准确率，可靠性较高；能够高速处理大量原始信号数据，有着较高的效率；同时由于采用预测方式进行异常检测，能够做到真正的实时甚至超前的关键参数预测分析。且本次研究提出并实现的异常检测方法与模型的适用范围不仅仅是本次实验所用的电源信号数据，而是各种场景下的频域数据均能够适用。因此，本次研究提出并实现的异常检测方法与模型具有高可靠、高效率、实时性、泛化能力强等优点。

（**上一段添加对比实验叙述？**）

### 1.3 论文章节安排

根据研究内容，将论文分为六个章节，其结构安排如下。

第一章为绪论。本章节首先介绍了本次研究机载电源信号频域品质参数异常检测的项目背景，介绍了我国自主研发C919大型民航客机的重要战略地位，接着介绍了大数据与机器学习相关的概念与技术特点，明确了本次研究的技术手段与主要目标。接下来，通过阐述当前电源品质参数异常检测研究方法的现状及其不足之处，引出本次研究的主要工作内容与研究意义，以及本次研究的贡献点。最后，阐述了本文的行文逻辑和组织结构，为下文研究问题求解做好铺垫。

第二章为本次研究设计模型的理论基础介绍。按照设计模型的结构顺序，首先介绍了时域频域转换相关的理论基础，包括快速傅里叶变换与采样定理两部分。接下来介绍了基础的循环神经网络原理与结构，然后引出对其变种优化结构长短期记忆模型（LSTM）的阐述，详细介绍了LSTM的内部单元结构与工作原理，以及其关键参数代表的含义。

第三章为本次研究提出模型的设计过程与设计思想。本章节首先介绍了电源信号频域品质参数的评价标准，然后详细介绍了当前广泛使用的阈值检测法，分析了其不能实时分析的不足之处。接下来从阈值检测法的这一痛点出发，以北京空气质量参数预测课题为例，详细阐述了基于LSTM的时间序列预测方法及其原理。然后顺延以上的分析思路，按照模型的结构顺序，详细介绍了本次研究提出模型的设计思想与该模型的具体结构。

第四章为对本次研究设计实现模型的实验验证及评估。本章节按照模型结构顺序，依次给出实验结果并作分析，并对最终的电源信号频域品质参数预测结果进行验证分析，评估本次研究设计实现模型的性能。

第五章为对比实验结果与分析。

（**待补充**）

第六章为结束语的总结与展望。总结了全文的主要的研究内容与创新点，并为后续深入研究提供思路与方向。

## 第二章 模型理论基础

### **2.1 时域-频域转换**

#### **2.1.1 快速傅里叶变换**

傅里叶变换是研究频域问题的核心，通过它能把信号从时间域变换到频率域，进而研究信号的频谱结构与变化规律。对于数字信号而言，离散傅里叶变换（Discrete Fourier Transform， DFT）是数字信号分析的基本方法，是数字信号分析与处理中的一种重要变换。

有限长序列x(n)的N点DFT为：

（**DFT公式**）

考虑x(n)为复数序列的一般情况，对某个值k，直接按照**上式**计算X(k)的一个值需要N次复数乘法与N-1次复数加法，因此，计算X(k)的所有N个值，一共需要N^2次的复数乘法运算与(N-1)^2次的复数加法运算，当N的值足够大时，(N-1)^2可以近似为N^2，可见，DFT的乘法与加法运算次数均为N^2。当N值较大时，DFT的总运算量将会相当大，这会极大地增加数字信号处理与分析的耗时。

在这种情形下，快速傅里叶变换（Fast Fourier Transform， FFT）算法应运而生，本文主要介绍基2FFT算法。基2FFT算法主要分为两类：时域抽取法FFT（Decimation-In-Time FFT，简称为DIT-FFT）与频域抽取法FFT（Decimation-In-Frequency FFT，简称为DIF-FFT）。

#### 2.1.2 采样定理

### 2.2 循环神经网络RNN

LSTM（Long Short-Term Memory）是循环神经网络（RNN）的一种重要结构。循环神经网络（RNN）源自于1982年由Saratha Sathasivam 提出的霍普菲尔德网络。循环神经网络能够挖掘出隐藏在数据中的时序信息，这是因为从网络结构上来说，循环神经网络能够记忆之前数据中的信息，并用这些之前的信息来影响后面时刻的结点输出。循环神经网络每个隐藏层的输入，不仅仅包含当前时刻的输入，其实还包含上一个时刻的输出。正因如此，循环神经网络适合处理时间序列的数据，它在时序分析、语音识别、自然语言处理等方面都有着出色的表现。

（**扩充叙述RNN**）

图2左边是RNN模型没有按时间展开的图，右边是按时间展开的RNN网络。

（**RNN结构图**）

但是，循环神经网络存在一个缺陷，那就是它可以保存的历史信息是有限的，换句话说，它无法记住时间序列中很长时间之前的信息，无法利用这些很久之前但依旧重要有效的信息去帮助给出当前的决策输出。LSTM网络的设计就是为了解决这一问题。正因为LSTM能够记住很长时间之前的信息，所以在很多时间序列任务上，LSTM的表现比循环神经网络要更好。

### 2.3 LSTM神经网络

#### 2.3.1 LSTM神经网络结构

LSTM网络结构最早由Sepp Hochreiter 和Jürgen Schmidhuber 于1997年提出。LSTM是一个拥有三个“门”结构的特殊网络结构，这三个“门”分别是输入门、遗忘门和输出门。LSTM通过“门”来控制丢弃或者增加信息，从而实现遗忘或者记忆的功能。这里的“门”是一种使信息选择性通过的结构，由一个sigmoid函数和一个点乘操作组成。因为使用sigmoid函数作为激活函数，所以“门”的输出值在[0,1]区间，输出值为0代表信息完全丢弃，输出值为1代表信息全部通过。除了输入和输出之外，LSTM网络结构还会记录每一个时刻过后的状态信息。

LSTM的基本单元结构如图3所示

（**LSTM结构图**）

遗忘门的输入有两个，分别是上一时刻的输出ht-1和当前时刻的输入xt，经过sigmoid函数激活后，遗忘门的输出值f与上一时刻的状态向量ct-1进行点乘操作，来控制上一时刻状态的遗忘程度，最终得到一个被过滤后的上一时刻状态向量f\*ct-1。

输入门的输入也有两个，一个是上一时刻的输出ht-1，另一个是当前时刻的输入xt。输入门有两个激活函数，一个是sigmoid函数，另一个是tanh函数。Tanh函数控制当前时刻的输入信息中有哪些可以被通过，输出为z，sigmoid函数控制这些被允许通过的信息的具体通过量，输出为i。i和z进行点乘操作后，再加上遗忘门的输出f\*ct-1，就生成了当前时刻的新状态ct。

（**补充sigmoid与tanh函数曲线图**）

输出门的作用是控制当前时刻的新状态ct中有多少信息被过滤掉。和输入门类似，它同样通过sigmoid函数和tanh函数来过滤状态信息。最终生成当前时刻的新输出ht。

（**LSTM各个门计算公式**）

以上就是LSTM网络的基本结构和工作原理，本次研究中采用上述的标准LSTM网络结构。

#### 2.3.2 timesteps参数

timesteps是LSTM网络模型中一个很重要的概念，它对LSTM网络的性能提升有着重要的影响。LSTM网络模型的输入一般是三维的：[采样数，timesteps，特征数]。在现有的LSTM模型中，timesteps的值一般默认取1。这种情况下，每一个时刻的输入数据都有一个输出。然而，这并没有把LSTM网络的优势体现出来。因为timesteps值等于1的时候，当前时刻对应的输出几乎只用到了当前时刻输入数据的信息，之前时刻的历史信息并没有很好地为给出当前时刻的输出而做贡献。LSTM网络的优势正是在于通过三个特殊的“门”结构去选择性地保留之前的之前时刻的历史信息，因此需要增大timesteps的值。

比如，假定timesteps的值为10，那么连续的10个时刻对应的输入数据可以看做一个长度为10的序列，前9个时刻的输入进入LSTM网络模型后，并不会给出最终输出，而是更新内部的记忆状态，直到第10个时刻，才会综合前9个时刻的数据信息给出最终的输出，这个最终输出是第10个时刻对应的输出。这种情况下，就可以发挥出LSTM网络的优势，使最终输出更加精确。

如图所示。左边是默认的LSTM模型单元的等效结构图，右边的是调整timesteps参数后的LSTM模型单元的等效结构图。可以明显地看出，相比于调整之前的模型，调整后的模型能够有效地利用更多来自之前数据序列中的历史信息。这些历史信息能够对LSTM模型的训练和预测起到重要的影响。因此，调整后的模型有着远高于调整之前模型的预测准确率。

因为原始数据是二维的，所以输入数据需要经过重塑处理，才能进入LSTM模型训练。当timesteps的值大于1时，数据还需要进行重叠操作。重叠的序列长度等于timesteps的值。这是因为每个时刻都需要给出一个最终输出，如果不进行数据重叠，会造成每隔timesteps个时刻才有一个输出，而在间隔内的时刻没有输出的问题发生。

/ \* 本次研究中使用的timesteps值为10，这个数值是综合考虑数据的采样间隔时间和模型训练时间之后确定的。\*/

### 2.4 其他变种RNN网络

### 2.5 本章小结

## 第三章 电源频域数据异常检测方法的研究设计

### 3.1 研究现状分析

#### 3.1.1 电源信号频域参数与评价标准

在飞机三相交流发电机电源信号频域的异常检测中，交流电源信号的基波频率与谐波频率是主要关注的两个参数指标，接下来对基波与谐波作简单介绍。在复杂的周期性振荡中，信号包含着基波与谐波。和该振荡最长周期相等的正弦波分量被称为基波，其频率值被称为基波频率。频率等于基波频率整数倍的信号中正弦波分量被称为谐波，谐波对应的频率统称为谐波频率。换言之，将非正弦周期性信号按傅里叶级数展开，得到的展开式中频率与原信号频率相同的量就是基波，展开式中频率为基波频率整数倍的剩余分量均为原始信号的整次谐波。

（**附傅里叶展开公式，与基波谐波的分解图像**）

谐波实际上是理想周期性信号上的附加噪声，它的产生主要由发电机自身产生、输配电过程产生、电力设备产生等原因导致。本次研究的对象是飞机三相交流发电机输出的电源信号，其中的谐波属于电源端产生的谐波。在飞机三相交流发电机设备中，受制作工艺影响，发电机的三相绕组在制作上很难做到绝对对称，且其铁心也很难做到绝对的均匀一致，因此发电机输出的电源信号中存在谐波是不可避免的。与此同时，谐波也有着诸多危害需要我们关注与防范。谐波会导致发电机存在附加损耗从而影响效率，且若谐波含量过高，极易引发从电源端到输配电再到用电设备的全线安全问题，小到信号干扰、计量误差，大到短路爆炸都是可能存在的。本次研究的背景是商飞C919大飞机的机载发电机电源，更是不容许出现任何安全隐患。谐波尚且如此，更不用提作为主频率输出的基波频率，对其的监控检测更是重中之重。正因如此，找到一种有效可靠的飞机发电机电源频域数据异常检测的方法就显得尤为重要。

本次研究中使用的原始飞机电源数据即为复杂周期性振荡信号，其中包含着基波与谐波，在业界有对飞机电源的评价标准，这个标准具有通用性与代表性，接下来给出具体标准数值。 基波对应的是飞机交流电源信号的主频率，基波的标准值为400Hz，合格范围是400 ± 1 Hz，超出该范围即可视为基波频率出现异常。谐波是所有整次谐波的统称，对单次谐波分量，要求其含量小于基波的4%含量，超出该含量即可视为该谐波分量频率出现异常。

#### 3.1.2 电源信号频域数据的阈值检测法

正如本文第一章引言中所叙述的，当前对电源信号频域数据的异常检测一般是使用传统的阈值检测法。阈值检测法是先将电源信号的时域数据进行快速傅里叶变换，得到电源信号的频谱，然后根据本文上一小节提到的电源参数评价标准来分析频谱，计算频谱中的基波与谐波是否超出阈值，也就是评价标准中的合格范围，从而判断电源的基波与谐波含量是否异常。阈值检测法的优点是简单易行，检测结果百分之百准确，不会出现误报故障的问题，这是因为它是基于实际的信号数据分析得出的，只要阈值给定，检测的结果就一定是准确的。

但是阈值检测法也存在着一些较为严重的问题与不足之处。比如，值得注意的是，这里方法中用到的电源信号时域数据是单位时间的数据量，可以是1秒的数据量，也可以是0.5秒或2秒的数据量。这里就会存在一个问题，检测的结果必然是迟滞的。因为阈值检测法中的快速傅里叶变换需要至少一个单位时间的信号数据量，即使我们忽略后续采集到的时域信号传输时间和频谱阈值分析计算时间，也存在着至少一个单位时间的延迟，这一个单位时间的延迟还仅仅是为了采集阈值检测法所需要的输入数据，并不包含数据传输和阈值检测法真正运行计算所需要的时间。如果考虑实际运行中的各项时间损耗，阈值检测法的检测结果可能会有较大的延迟。换言之，目前通用的阈值检测法的检测结果不具备实时性，如图所示，获取检测结果的时刻与对应的信号采集时刻之间存在延迟。

（**检测结果时刻与采集时刻延迟图像**）

这种现象存在着一定的安全隐患，因为如果延迟过大，阈值检测法分析的结果是没有意义的，就像马后炮一样，故障发生后才检测出来，这就失去了异常检测报警的意义。尽管阈值检测法的结果是百分百准确，但从算法设计目的出发，我们是为了进行异常检测，异常检测的目的是为了监控预警，提前预报可能出现异常的时刻以及故障位置才是我们希望看到的效果。更何况由上一小节可以知道本次研究的背景是飞机交流发电机电源信号频域数据异常检测，从安全的角度出发，应当做到实时或者提前预警预报。

阈值检测法不具备实时性，其检测结果虽然准确但参考意义不大，本次研究的目的也正是找到一种能够做到对机载电源信号频域数据进行实时检测或者提前预警的异常检测方法。想达到提前预警的效果，最好的办法就是对检测结果进行预测计算，如果预测的准确率能够满足需求，则可认为这种预测计算的方法是有效可行的。

#### 3.1.3 基于LSTM的时间序列预测方法

上一小节中提到的本次研究的目的找到一种适用于电源信号频域数据的预测计算方法。提到基于预测的异常检测方法，现有的许多方法都可以对时间序列数据进行有效的预测分析，如长短期记忆网络（LSTM）、差分整合移动平均自回归模型（ARIMA）、小波变换等等。这里本文主要介绍基于LSTM的时间序列预测方法。

（**根据第五章对比实验内容这里可增删**）

在介绍基于LSTM的时间序列预测方法之前，需要先介绍无监督学习与监督学习相关的内容。在机器学习中，按照标记的有无可以分为无监督学习和监督学习。无监督学习中使用的训练集与测试集是没有标记的，常见的无监督学习算法有主成分分析法、K-means聚类算法等等。监督学习中的训练集与测试集是有标记的，可以理解为模型有输出的期望值，并通过多轮训练模型不断调整模型内的结构参数，使模型输出能够最大程度地接近输出的期望值。应用广泛的神经网络与决策树都属于监督学习。

本次研究使用的LSTM网络模型属于神经网络中深度学习的扩展分支，同样在监督学习的范畴内，其训练接与测试集数据是需要具有标记项的，也就是需要有输出项。然而在直接获取的原始时间序列数据中，绝大多数是没有标记项的，需要根据研究内容人为地选择并添加标记项。以下以预测北京天气质量的案例为例进行说明。

（**北京天气质量原始数据表格，无输出，训练集**）

从上表可见，获取的训练集与测试集原始数据均没有标记项。该案例研究的内容是预测空气质量参数中的PM2.5与PM10两种参数，因此，需要将PM2.5与PM10这两个维度的数据提取出来作为标记输出，如下表所示。

（**北京天气质量原始数据表格，包含输出（未平移），训练集**）

只是提取目标维度数据列作为标记输出还不够，因为案例的目的是预测空气质量参数PM2.5与PM10，**（上表）**中的标记输出是每个采样时刻一一对应的实测参数值，而不是期望输出的预测值。由于期望的标记输出是预测性质的，是未来某个时刻的待输出参数值，且案例研究的对象是时间序列，因此可以使用未来某个采样时刻的真实测量记录值作为当前采样时刻的标记输出。将**（上表）**中的标记输出进行纵向向上的平移若干次，再舍去表格首尾包含缺失值的行记录数据，就可以得到适用的标记输出，如下表所示。

（**北京天气质量原始数据表格，包含输出（平移一次），训练集**）

（**北京天气质量原始数据表格，包含输出（平移三次），训练集**）

**上左表**中每个当前采样时刻对应的标记输出是未来下一个采样时刻的真实测量记录值，它同时也是当前采样时刻的预测期望值，表示当前采样时刻的输出是下一采样时刻PM2.5与PM10的预测值。这样就可以实现提前一个时间单位的预测。同理，**上右表**是提前三个时间单位预测的数据集，即当前采样时刻输出的是三个时间单位后PM2.5与PM10的预测值。通过这种方法就可以将原始数据集调整为适用于实现预测功能的数据集，且不改变数据集是时间序列的这一性质，代价仅仅是损失了首尾的几行数据而已，相对于训练集与测试集的庞大规模，这种程度的数据量损失几乎可以忽略不计。

由本文第二章中对于LSTM神经网络模型的原理介绍可知，LSTM神经网络模型非常适合用来处理时间序列数据。用LSTM神经网络模型对调整后的时间序列数据集进行多轮的训练与测试，理论上可以有较好的预测准确性。以上便是对基于LSTM的时间序列预测方法的简单介绍。

### 3.2 模型设计思想与实现

由本文第二章的理论基础部分可以知道，LSTM神经网络是十分适合用来处理时间序列的，而且上一小节介绍了基于LSTM的时间序列预测方法，可以通过预测的方式实现对时间序列数据的实时甚至预测性的分析。然而，本次研究的对象是电源信号的频域数据，并非时间序列数据，根本不具备时域中的一些特性，因此是无法直接用LSTM神经网络模型来处理的。

本文正是基于以上两个现状，思考并研究是否能够将LSTM神经网络模型利用在分析频域信号数据上，研究出一种新的具有**预警性、实时性、高可靠、高效率**的频域异常检测方法，来替代传统的频谱阈值检测法。

如上所述，LSTM神经网络模型仅适用于处理分析时域上的信号数据，对频域信号数据却无能为力，这是由于时域与频域的特性所决定的，频域信号数据映射到时域中对应的是完整的时域信号数据，相当于LSTM神经网络模型只有一行单独的输入数据，而不是连贯的时间序列信号数据，连训练网络模型都做不到，显然是不适用来获取理想预测结果的。

本次研究提出的基于时域-频域-时域转换与LSTM神经网络模型的电源信号频域数据异常检测模型如下图所示。

（**模型总体结构图**）

如图所示，模型主要分为时域-频域-时域转换、数据预处理和LSTM神经网络三个部分，其中数据预处理部分又由数据降维、数据清洗、数据缩放、数据标记四个部分组成。

#### 3.2.1 时域-频域-时域转换

本文提出一种新的思路来综合利用时域-频域-时域转换和LSTM神经网络模型。其核心思想是对电源信号的电压-时间数据（V-t）进行分段，分为一个个单位时间（如1秒）的时域时间子序列片段。然后对这些时间序列片段分别进行快速傅里叶变换，获取这些子序列片段对应的频域信号数据（F）。再按原先的时序依次将这些频域信号数据串联，来生成一个完整的频率-时间数据（F-t），这个频率-时间数据在时域的表现上是离散而非连续的，但每个离散的单位数据之间却具有时序上的连贯性，是按时序排列的。

（**如图所示，时频时转换过程示意图**）。

通过这样一个拆分后再组合的过程，完成一个从时域到频域，然后再到时域的转换，从而生成可以被LSTM神经网络模型使用的时间序列数据。这个时间序列数据实际上是由一系列时间上连贯的频域数据拼接而成的，可以理解为做了一个降采样率的操作，用高采样率的原始数据生成了低采样率的时域数据，同时这个时域数据也保留了原始数据的频域信息。这样就可以使用LSTM神经网络模型来对频域中的信息进行处理与分析预测。

对原始的飞机交流发电机电源信号数据，先进行分段的时频转换，然后将频域按时序串联得到完整的频谱数据。这样得到的频谱数据还不能直接用于LSTM神经网络模型进行训练，需要进行一系列预处理操作。

#### 3.2.2 数据预处理

对于机器学习任务来说，数据预处理是十分重要的，它在很大程度上影响了模型最后的结果。对于本次处理电源信号数据的任务来说，数据预处理的流程可分为：数据降维、数据清洗、数据缩放和数据标记四个部分。

首先是数据降维。数据降维通常又被称为特征提取，在很多数据处理的任务中，需要被处理的数据都属于高维数据，具有非常多的特征数目。 如果不进行降维处理，直接进行分析通常会带来极大的计算量，一方面这会使后续的模型训练时间大大增加，另一方面，过多的特征数目会带来很大的噪声，影响模型的收敛和最终结果。 因此，在直接对数据进行分析之前需要对先提取出高维数据的特征，在保证结果不受影响的情况下，尽可能地减小原始数据的维数。本次实验中通过时域-频域-时域转换得到的完整频谱数据量是十分庞大的，这是因为转换后的频谱数据维度太高，且其中的绝大多数特征分量是我们不需要关心的。因此需要对完整频谱数据进行降维处理。

一般来说，降维算法有主成分分析法（PCA）、因子分析法（FA）、独立成分分析法（ICA）、奇异值分解法等等。以上降维算法适用于特征分量之间的关联性不易被显性发现的情形，其中最常用的是主成分分析法。但是本次研究中使用的频谱数据较为特殊，由于研究目的是进行基波与谐波的异常检测，因此不需要使用上述这些具有普适性的降维算法，而是采用人工剔除筛选特征的方法来降维。这里我们只需要关注电源信号的基波与主要高次谐波即可，其中次数过高的谐波由于其含量过低可以被直接剔除，而且由于本次研究采用的是平衡三相发电机，所以偶次谐波也忽略并剔除，只保留基波和若干低次的奇次谐波即可。

（**附降维前后对比图**）

然后是数据清洗。对降维后的频谱数据进行检查和清洗是十分重要的，数据清洗分为去除重复数据和填补缺失值两部分。对于本次采用的电源信号频谱的时序串联数据，去除重复数据可以检查时间序列，以时间序列为索引，去除采样时刻相同的重复数据。然后检查原始数据中是否有缺失值，如果有，就采用k最近邻算法进行缺失值的填补。这里对于k最近邻算法不再重点介绍。

接下来是数据缩放。本次研究使用的电源信号频谱数据的数值变化区间较大，既存在较大数量级的频率幅度数值，也存在较小数量级的频率幅度数值。因此需要将频谱数据缩放到指定的区间中，以增强数据的稳定性，同时也便于后续模型的计算处理。本次研究将电源信号的频谱数据缩放到0和1之间。

最后是数据标记。这一步的作用是规范LSTM神经网络模型的输入输出维度，以及生成符合LSTM神经网络模型要求的输入数据。机器学习根据标记信息的有无，可以大致分为监督学习和无监督学习两类。其中监督学习可以分为分类和回归两类问题。本次研究的内容是如何对飞机交流发电机电源数据的频率参数进行有效预测，来完成实时的异常数据检测，这实际上属于回归问题。然而，本次研究用到的电源信号频谱数据是没有标记信息的，无法直接用于模型训练，所以频谱数据应当先被标记。对于本次研究采用的电源信号频谱数据来说，由于频谱数据整体上是以时间序列为索引的，所以只需要把待预测的频率参数值序列前移若干个时间标度，然后得到的新序列就是数据集的标记信息。这部分的详细原理可以参考本章3.1.3小节中对时间序列的实时预测部分的阐述。简单来说，每个时间标度的完整频谱数据就是LSTM神经网络模型的输入数据，对应的标签是未来若干个时间标度后的待预测频率参数数值，同时这部分也是LSTM神经网络模型的输出数据。

#### 3.2.3 LSTM神经网络模型

本文使用的LSTM神经网络模型结构如图所示

（**LSTM网络结构图**）。

经过预处理后的数据直接进入LSTM层，在LSTM层后面是若干全连接层与dropout层所组成的中间层，最后是一个输出维度为5的全连接层（该模型每次对五个频率分量数据进行预测，分别为基波与四个高次谐波）。实际上，如果把这些中间层移除，模型依然可以正常工作。然而，如果添加了这些中间层，模型的预测效果会更好。

中间五层全连接层的作用是增加网络模型容量，提升模型复杂度。从深度学习的理论上来说，参数越多的模型复杂度越高、容量越大，这种模型由于具备复杂的结构，经过多次的训练，可以生成相较于单层网络更加准确、更加契合任务目标的网络模型，简而言之就是可以完成更加复杂的机器学习任务。换句话说，添加这些全连接层可以显著增强模型的性能。增加神经网络的层数是最简单的深度学习方法。但是，全连接层数目并不是越多越好。过多的中间隐藏层数目会带来收敛困难、计算量过大等等问题。其中计算量过大这个问题在如今已经不能算作是值得关注的问题，由于现在硬件机器算力的飞速进展，现在的机器算力已经足够胜任深度学习复杂网络模型的学习训练任务。尽管如此，层数过多的深度学习网络模型收敛困难的问题依旧存在因此，因此选取合适的中间隐藏层数目十分重要，这对LSTM网络模型性能的提升有很大的帮助。

除了全连接层以外，中间层还包括dropout层。Dropout层的作用是防止网络模型训练出现过拟合的问题。由于本次研究设计的LSTM神经网络模型采用了深度学习的理论方法，有着较多的中间隐藏全连接层，因此模型的复杂度很高，模型内的参数也非常多，在这种情况下，如果训练使用的数据量过小的话，网络模型很容易出现过拟合的问题。过拟合问题可以理解为高射炮打蚊子——虽然不够准确，但确实是大材小用的感觉。用过小的数据量训练过于复杂的网络模型，导致性能强大的网络模型过拟合，在训练数据与验证数据上表现很好，预测准确率很高，但在测试数据上的预测结果很差，也就是泛化能力很差，无法普遍地应用在不同数据上，不具备实用性。这种结果显然是不能接受的，就需要用到dropout层来解决训练中网络模型过拟合的问题。

（**dropout图示结构，两个图，标准神经网络、临时删除节点网络**）

 

如图所示，dropout层的工作原理是根据设置的比例，在该层中随机临时删除掉给定比例的神经元，然后在临时删除神经元后的网络上进行后续训练步骤，具体是前向传播、反向传播与修改更新剩余神经元参数。之后就是重复该过程，先恢复之前临时删除的中间层神经元，然再随机临时删除并训练，对于训练后的神经元参数更新，未被选中临时删除的神经元进行参数更新，被选中临时删除的神经元不参与参数的更新。以上就是dropout层防止过拟合问题的工作原理，通过随机删除中间层的部分神经元，来降低网络模型与训练数据之前过高的契合度，增强模型的泛化能力，以此来防止过拟合问题。因此，在本次研究中的LSTM神经网络模型中设计增加若干dropout层以防止过拟合。

### 3.3 本章小结

## 第四章 **模型实验结果分析**

上一章设计实现了电源信号频域数据的处理预测模型，本章对该模型进行实验验证，并分析该模型各个部分的实验结果及最终的预测结果。通过实验，阐述本次研究设计实现的电源信号频域数据处理预测模型的可行性以及准确性，验证该预测模型的高可靠、高效率以及可推广应用的优点。本次实验的环境搭建如下表所示。

（**环境配置表格，语言、硬件配置等**）

（**提前1秒，2秒，5秒预测对比实验，结果发散推广，不仅仅适用于飞机电源，也适用其他频谱场景**）

### 4.1 实验原始数据介绍

（**下采样不影响结果。说明原因。**）

本次研究使用的飞机发电机电源信号数据来源于研究背景项目中的真实数据，由商飞公司提供，数据内容如下图所示。



如图所示，原始数据包含机载四个三相发电机的电源信号数据，每个电机的数据又分为三个相的电压与电流数据。为了简化实验流程，本次研究只选取1号发电机A相电压数据作为原始电源信号数据，即本次实验只对1号发电机A相电压信号的频域参数进行预测分析。由于研究实验方法相同，设计实现的模型与实验方法也同样适用于其他相位的电压、电流信号数据，因此这里可以简化实验流程，以1号发电机A相电压数据的预测分析结果为代表，来评估本次研究设计实现的方法与模型。

选取原始电源信号数据V1\_A后，还需要进行降采样处理原始数据。由于商飞公司提供的原始数据量过小，进行时-频-时转换后生成的时间序列过短，无法用于LSTM神经网络模型训练出有效可靠的模型参数集合，因此，需要将原始电源信号数据进行降采样再拼接的扩充序列长度处理。原始的飞机发电机电源信号具有很高的采样率，即使降采样之后，根据采样定理，仍然足够支持获取实验需要的高次谐波信息。

（**降采样再拼接的过程图？有必要说明？**）

此外，降采样的操作并不会影响实验的预测分析结果，因为降采样的原因仅仅是因为原始数据时间序列长度过短从而无法用于训练，是受原始数据提供方商飞公司的限制，为了正常完成实验而执行的操作。其并没有对本次实验设计实现的预测分析模型产生任何影响与改变，如果本次实验的结果经评估验证是达到预期且有效可靠的话，即使更换为足够数量的原始高采样率数据，采用本次研究设计实现的预测分析模型，重新实验训练模型、调整参数，也一样可以获得期望的高准确度预测效果，而且结果应该会更好。遗憾的是受限于数据提供方的限制，本次实验不得不对原始数据略作修改以验证本次研究的方法与模型，但不会对结果的正确性有任何影响。

修改调整后的原始数据形式如下表所示。

（**采样率2500Hz 长度560 原始数据部分行记录表**）

如表所示，原始数据是单维度的电压-时间（V-t）电源信号数据，采样率fs为2500Hz，由采样定理可知，将原始电源信号数据进行傅里叶变换后可获取的最高频谱频率fmax为采样率fs的一半，即为1250Hz。（**2500/2 = 1250， 1250\*4 = 5000**）又因为原始电源信号数据是经过降采样再拼接的修改调整，因此实际的最高频谱频率是四倍的fmax，即5000Hz。根据行业标准可知，飞机发电机电源信号的基波频率标准值为400Hz，因此使用该原始数据可以最高分析到12次谐波，即4800Hz的频谱信息，理论上来说是足够的。

原始电源信号数据时域电压-时间（V-t）部分波形图如下。

（**原始电源信号数据时域电压-时间（V-t）部分波形图**）

### 4.2 时域-频域-时域转换结果

根据本文第三章设计实现的电源信号数据预测分析模型，实验的第一步是对原始电源信号数据进行时-频-时转换。

本次实验将原始电源信号数据分为560个分段，每个分段对应时长1秒的采样数据，这里的1秒也同样是时-频-时转换中的单位时间长度。下图是第一个分段的时域波形图。

（**1/560， 第一个分段的波形图**）

接下来对每个分段分别进行快速傅里叶变换，得到对应的频域数据集合，同样是560个。下图是第一个分段的频谱图像。

（**1/560， 第一个分段的频谱图像**）

将560个分段全部转换到频域后，按照分段之间的时序关系，把560个频域分段依次串联拼接，得到时-频-时转换部分的处理结果，即该部分输出的时频（F-t）数据。下图为时频图的部分图像内容。

（**时频图（部分）**）

### 4.3 数据预处理

实验的第二步是数据预处理部分，该部分由三个子步骤组成，依次是降维处理、数据标记、数据缩放。

根据本文第三章3.2.2小节中的阐述，本次实验采用人工选取特征的方式进行降维处理。由本文4.2小节的分段频谱图图像可以看出，三相发电机电源信号的频谱中只有基波与部分奇次谐波的幅值是显著的，其余如偶次谐波等频率分量的幅值非常低，可以忽略不计。这一实验结果也印证支持了本文第三章3.2.2小节中的理论分析结果。本次实验选取的频率分量如下表所示。

（**选择的频率分量表，多维度，一行**）

本次实验所选的频率分量是基波与其三次、五次、七次、十一次谐波，同时也选取了它们分别左右最近邻的频率分量值用于训练LSTM神经网络模型，这些频率分量均为电源信号频域数据异常检测需要重点关注的频率分量。降维处理后的部分时频数据如下表所示。

（**降维处理后的数据表格（无标记）**）

数据预处理的第二个子步骤是数据标记。根据实验的异常检测分析目标，选取合适的模型输出。这里和降维处理选取频率分量的依据相同，同样选择基波与其三次、五次、七次、十一次谐波作为LSTM网络模型的输出，将选取的五个维度的数据列提取并根据本文第三章中3.1.3中介绍的方法添加标记。此处标记的超前延时为一个单位时间，即一秒钟。添加标记后的数据部分内容如下表所示。

（**添加标记后的数据表格（超前1秒）**）

数据预处理的最后一个步骤是数据缩放。本次实验采用Python中sklearn集合包中的MinMaxScaler工具类对数据进行缩放，类似于软件中的编解码。这里先将数据缩至[0,1]区间内，以便于后续模型参数的计算。调整后的数据部分内容如下表所示。

（**数据缩放后的表格**）

### 4.4 模型的训练与预测结果分析

数据预处理结束后，就可以开始本次实验的最后一步，进行LSTM神经网络模型的训练与参数调整。首先将预处理后的数据划分为训练集、测试集与验证集。根据本文第二章2.3.2小节的阐述，实验需要选择合适的timesteps步长参数值。本次实验选取的timesteps步长参数值为10，因此还需要对训练集与测试集的数据再进行维度调整，才可以作为LSTM神经网络模型最终的输入数据。预处理后的数据是二维数据，形式为[time, var]，第一个维度是采样时间，第二个维度是选取的频率分量与输出。LSTM神经网络模型输入数据需要三维数据，形式为[time, timesteps, var]，因此需要增加维度，根据设置的timesteps步长参数值调整训练集与测试集的数据形式。

（**附代码片段？**）

然后是LSTM神经网络模型的训练阶段，经过调参与多次的训练，确定最终训练参数如下表所示。

（**训练参数表，timesteps步长、轮数、batch\_size、损失函数、lstm内部单元数等等**）

训练结果如图所示。

（**训练结果图像，训练集与测试集**）

由上图可知，该参数配置下的模型训练结果已经收敛，且误差较低，可以用于在验证集数据上进行验证分析。验证预测结果的均方根误差（RMSE）如下表所示，需要注意的是，在分析预测数据之前，已经将预测数据通过MinMaxScaler工具类恢复至原始数据的大小范围，因此RMSE值可能会略微偏大。

（**各频率分量的预测RMSE结果表格**）

LSTM神经网络模型输出的五个频率分量预测图像如下一系列图像所示。

（**五个频率分量的实际与预测对比图像**）

预测结果的准确率与程序运行时间如下表所示。

（**预测结果准确率与运行时间表格**）

由以上预测结果图表可以看出，本次实验的预测结果能够达到期望的效果，对电源信号频域数据异常的检测准确率均在95%以上，对输出的各个频率分量的幅值预测也能够做到跟随实际幅值的波动，做出同样波动趋势的有效预测。可以看出，本次研究设计实现的模型有着较强的学习能力，训练后模型的预测值能够准确地跟随实际值的变化进行变化，电源信号频域参数的预测值与实际值的图像拟合得很好。

本次实验结果表明，本次研究设计的飞机发电机电源信号频域数据异常检测方法与预测分析模型，能够实时或者超前地预警飞机发电机电源可能发生的故障，且虚警率低于5%，运行时间少于1秒，达到了预期的预警性、实时性、高效率、高可靠的性能要求。

（**与阈值检测法结果对比**）

### 4.5 调参实验结果与分析

改变超前秒数、改变timesteps、改变损失函数mse/mae等等。

### 4.6 本章小结

## 第五章 对比实验结果与分析

短时FFT？小波变换

## 第六章 总结与展望