文章编号:1007 - 4252(2023)02 - 0002 - 117 DOI:10.20027/j.gncq.2023.0018

基于神经网络的硅晶圆表面形貌快速量测方法研究

黄嘉晔,关子钧,李卫民
(上海集成电路材料研究院有限公司,上海 201899)

摘要:硅晶圆平坦化抛光后的表面形貌量测可以定量化晶圆平坦度,是品控的关键一环。然而,高平坦度硅晶圆表面检测设备原理复杂,检测位点过万,时间长达 1 分钟,同时设备体积大,难以集成在抛光机台中对硅片进行原位检测,这限制了生产效率。机器学习技术数据分析和预测能力优异,应用工业生产领域,能够提高生产效率和精度。本研究针对 300mm 硅片基于集成电路材料基因组平台开发了一种基于神经网络的晶圆形貌预测模型,该模型极大地减少表征晶圆形貌所需要的输入信息,通过 14 个关键点的厚度信息,便可得到绝对平均误差为 4.07nm 的预测值,从而仅需 10s 左右的时间对晶圆形貌进行高精度的预测,降低了量测设备的硬件要求,为晶圆平坦化工艺中的快速量测提供思路。

关键词:晶圆表面形貌;快速量测;神经网络
中图分类号:TN407 **文献标志码**:A

A neural network-based method for rapid measurement of silicon wafer surface topography

HUANG Jiaye, GUAN Zijun, LI Weimin
(Shanghai Institute of IC Materials, Shanghai 201899, China)

Abstract: Surface profile measurement of silicon wafers after flattening and polishing can quantify wafer flatness and is a key part of quality control. However, the principle of high flatness silicon wafer surface inspection equipment is complex, with over 10,000 inspection sites and time up to 1 minute, while the large size of the equipment makes it difficult to integrate in the polishing machine for in-situ inspection of silicon wafers, which limits the production efficiency. Machine learning techniques with excellent data analysis and prediction capabilities can improve production efficiency and accuracy when applied in industrial production. In this study, a neural network-based wafer profile prediction model is developed for 300 mm wafers based on the IC material genome platform. The model greatly reduces the input information required to characterize the wafer profile, and a prediction value with an absolute average error of 4.07 nm can be obtained from the thickness information of only 14 key points, resulting in a highly accurate prediction of the wafer profile in only about 10 s and reduce the hardware requirements of the meas-

收稿日期:2023-03-20; **修订日期**:2023-04-21
基金项目:上海市集成电路科技支撑专项:基于国产材料的 300mm 大硅片化学机械抛光工艺研究“21DZ1101200”
作者简介:黄嘉晔(1983-),男,博士,工程师,主要研究方向为集成电路工艺与材料(E-mail:jiaye.huang@sicm.com.cn).

urement equipment. The prediction of the wafer shape can be performed with high accuracy in only 10s, providing an idea for fast measurement in wafer flattening process.

Key words: Wafer Surface Topography; Rapid Measurement; Neural Network

0 引言

先进制程芯片对硅片表面平坦度要求苛刻,比如 14nm 制程用硅片表面在任意 26mm×8mm 面积内平整度≤20nm。在硅片成型工艺中,硅晶圆平坦化抛光工艺后,需要对表面进行形貌检测。

传统的晶圆抛光后的形貌测量方法可以分为两类:用于实验中的形貌测量方法和工业生产中的形貌检测方法。实验中晶圆抛光后的形貌量测方法主要有扫描电子显微镜、光学显微镜、三维激光测量仪、投影测量仪、接触式触探等。在工业生产中,晶圆形貌的量测需要达到极高的灵敏度和空间分辨率[1]和更紧凑的设备设计等要求,实验中采用的测量方法并不适用。

工业生产中以美国 KLA 公司高灵敏度和空间分辨率的干涉测量方法[2]为主流量测方法。

KLA 公司的 WaferSight2+是商用中最精确、最被认可的高平坦度硅晶圆形貌度量设备。设备采用双侧 fizeau 干涉仪同时进行硅片正反面测量,通过将反射的光束成像到相机上,获得晶圆的干涉图,每个相机像素处的干涉条纹相位经过高性能相位算法转换为晶圆表面高度,以达到亚纳米级别的测量精度,满足 15nm(SFQR)的平整度要求。硅片表面形貌的较大翘曲往往超出 fizeau 干涉仪的单次测量动态范围,Wafersight2+以孔径拼接技术[4]解决该问题,并通过数据拼接实现全部形貌的检测,扩大了动态测试范围,却使得整体量测元件尺寸很大,同时光路系统复杂,不具备集成于平坦化机台的可行性[3]。在硅晶圆平坦化工艺中,无法原位检测硅晶圆表面形貌,不能实现对其前馈和反馈的控制,且单片硅晶圆检测耗时长达 1 分钟,限制了生产效率及成本。目前,硅晶圆市场庞大,国产化硅晶圆快速量测设备处于空白阶段,晶圆表面形貌快速量测技术有望突破上述限制,扩大生产经济效益。

近年来,机器学习技术不断发展,在硅晶圆检测的应用中具有突出表现。王晓茹对基于机器学习的硅片图像缺陷检测算法进行了研究,提高了在线硅片表面缺陷检测精度和效率[5]。王梓铭基于机器视觉技术对硅片进行检测,并采用支持向量机算法对硅片进行分类,探索出了一种有效的硅片良品检测方法[6]。考虑到晶圆表面形貌尺寸特征小,检测需要采集的信息量大,数据特征复杂的特点,本文借助了神经网络算法。神经网络算法是一种模仿人脑神经元结构的算法,它可以用来解决复杂的分类和回归问题。它的基本原理是将输入数据映射到一组权重,然后通过一系列的计算,将输入数据转换为输出结果。神经网络算法的优势在于它可以从大量的输入数据中学习分析特征,并且根据输入数据的不同而产生不同的输出结果,以此从大量的晶圆形貌检测点中筛选出关键点。因此,检测设备只需检测少量的关键点信息便能预测出高精度的晶圆表面形貌,降低了检测设备的检测时间、量测设备的硬件要求,使晶圆平坦化的快速量测具备可行性。

本研究利用 Pytorch 工具包,建立了基于神经网络的机器学习模型用于曲面形貌预测。模型主要利用多层感知器(MLP, Multilayer Perceptron)从现硅片的高度数据中筛选关键点坐标,挖掘关键点高度对曲面分布规律的影响。利用学习到的规律,模型能够基于待测试硅片曲面上的少量关键点高度对完整曲面形貌进行预测。此外,模型能够基于预测结果筛选差异较大的部分作为新加入的特征点进行迭代以逐步优化对曲面的拟合效果。

1 正文

1.1 数据采集与建模方法

以上海集成电路材料研究院的集成电路材料基因组平台作为本研究中建立机器学习模型所使用数据集来源。该平台从产业界实际需求出发,运用材料跨尺度计算,结合研发过程中积累的实验数据,形成集成电路材料基因库,再利用机器学习的手段进行分析建模,与材料跨尺度计算相互协同,探索材料构效关系模型,加速集成电路中“卡脖子”问题的关键突破,其方法模式如图 1。

该数据库包含产业界材料性能数据、材料工艺数据、材料验证数据等。从中获取了国产硅片龙头产商的 250 片终抛后硅晶圆测试数据作为模型的数

据集。硅晶圆形貌检测设备为美国 KLA 公司的 WaferSight2+,保证了模型数据可靠性。使用 python 作为建立机器学习模型的编程语言,并利用 Pytorch 工具包搭建神经网络框架。

图 1　集成电路材料基因组平台

Fig. 1 Integrated Circuit Materials Genome Platform

1.2　神经网络算法原理

神经网络算法基于人工神经元模型和神经元间的连接权重,通过训练数据自适应地调整神经元间的连接权重,从而实现对输入数据的分类或预测[7]。

神经网络的基本构成包括输入层、隐藏层、输出层及信号传递的基本单元神经元(如图 2)。输入层用来接收外界的输入信号,输出层用来输出神经网络的结果,而隐藏层在输入层和输出层之间进行信息的传递和处理。

每个神经元都会收到来自其他神经元的输入信号,并将这些输入信号与自身的权重相乘后,再将它们相加起来,并将结果通过激活函数 S 进行非线性变换,将限定神经元输出在特定范围内,以便更好地适应实际问题的特征。

神经网络的训练过程是通过反向传播算法来调整神经元之间的连接权重,使得神经网络的输出结果与训练数据的真实结果相一致。

在使用反向传播算法调节前,神经网络将分两步正向计算各层的输出信号:(1)利用连接权重与神经元的输入信号进行组合;(2)应用激活函数处理组合后的信号,生成该层的输出信号。用矩阵运算表示为:

$$\vec{X} = \vec{W} \cdot \vec{I} \tag{1}$$

$$\vec{O} = \text{Sigmoid}\ (\vec{X}) \tag{2}$$

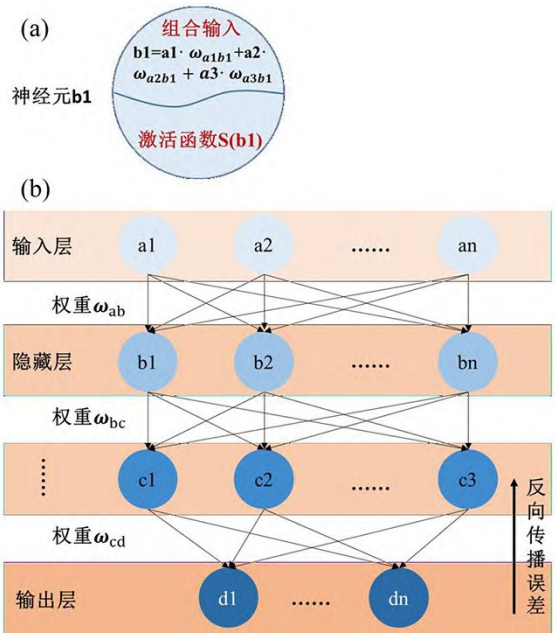


图 2　神经网络算法原理示意图

Fig. 2 Neural network algorithm principle schematic

其中 $\vec{X}$ 是组合后的信号矢量,$\vec{W}$ 是该层的各神经元与前一层的各神经元之间的连接权重矩阵,$\vec{I}$ 是前一层的输入信号矢量;$\vec{O}$ 是该层的输出信号矢量,Sigmoid 是所采用的激活函数 S。

接着,神经网络通过输出结果进行误差反向传播,这个过程将隐藏层中每一个神经元的误差定义为与该神经元正向连接中所有连接权重分割的误差之和。隐藏层神经元的误差矩阵形式为:

$$\begin{bmatrix} e_{\text{隐藏},1} \\ e_{\text{隐藏},2} \end{bmatrix} = \begin{bmatrix} \dfrac{w_{1,1}}{w_{1,1}+w_{2,1}} & \dfrac{w_{1,2}}{w_{1,2}+w_{2,2}} \\ \dfrac{w_{2,1}}{w_{1,1}+w_{2,1}} & \dfrac{w_{2,2}}{w_{1,2}+w_{2,2}} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{3}$$

权重矩阵中每个元素的分母是一种归一化因子。可将简化为:

$$\begin{bmatrix} e_{\text{隐藏},1} \\ e_{\text{隐藏},2} \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \tag{4}$$

其中的权重矩阵就是神经网络正向计算输出中公式(1)中的连接权重矩阵。因此,继续用矩阵运算来表示以上所描述的自输出层反向传播误差过程中计算得到的隐藏层误差值,则为:

$$\vec{E}_{\text{隐藏}} = \vec{W}_{\text{隐藏-输出}} \cdot \vec{E}_{\text{输出}} \tag{5}$$

其中,$\vec{E}_{\text{隐藏}}$ 是隐藏层神经元误差矢量,

$\vec{W}_{隐藏-输出}$ 是输出层与隐藏层之间的连接权重矩阵，$\vec{E}_{输出}$ 是输出层的误差矢量。

在神经网络的训练过程中，将定义一个损失函数来衡量神经网络输出结果与训练数据的真实结果之间的差异，并通过梯度下降算法来求解损失函数的最小值，以调整神经元之间的连接权重。进而逐步改善神经网络的输出值，直至与训练样本的真实输出值之间的误差达到设定范围内为止。

通过公式(5)可知，神经网络的误差是连接权重的函数。在最小化误差的过程中，使用梯度下降法来计算出误差函数相对于连接权重的斜率。所以，对于第 k 个神经元来说，与前一层中连接的第 j 个神经元的连接权重 $\vec{\omega}_{j,k}$ 的改变方向与斜率方向相反，其更新公式为：

$$\omega_{j,k}(new) = \omega_{j,k}(\text{old}) - \eta \times \frac{\partial E}{\partial \omega_{j,k}} \tag{6}$$

其中，η 是学习率，用来调节权重改变的强度。

对于第 k 个神经元，其输出值为 o_k，而真实的输出值为 t_k，所以，斜率公式为：

$$\frac{\partial E}{\partial w_{j,k}} = \frac{\partial}{\partial w_{j,k}}(t_k - o_k)^2 = -2(t_k - o_k) \times \frac{\partial o_k}{\partial w_{j,k}} = -2(t_k - o_k) \times \frac{\partial}{\partial w_{j,k}}\text{sigmoid}\left(\sum_j w_{j,k} \times o_j\right) \tag{7}$$

而 S 函数的求导公式为：

$$\frac{\partial \text{sigmoid}(x)}{\partial x} = \text{sigmoid}(x)(1 - \text{sigmoid}(x)) \tag{8}$$

因此，公式(7)可写为：

$$\frac{\partial E}{\partial w_{j,k}} = -(t_k - o_k) \times \text{sigmoid}\left(\sum_j w_{j,k} \times o_j\right)\left(1 - \text{sigmoid}\left(\sum_j w_{j,k} \times o_j\right)\right) \cdot o_j \tag{9}$$

用矩阵运算来表示公式(6)所描述的由误差函数的斜率矩阵来更新连接权重矩阵的话，则为

$$\vec{W}_{j,k}(new) - \vec{W}_{j,k}(old) = \alpha \times \vec{E}_k \cdot \vec{O}_k(1 - \vec{O}_k) \cdot \vec{O}_j^T \tag{10}$$

1.3 基于神经网络的快速量测模型搭建

模型利用的 MLP 为一个包含一层输入层 L_I，隐藏层 L_H 和输出层 L_O 的全连接神经网络，分别由 n_1, n_2, n_3 个神经元构成。其中 L_I 与 L_H 之间、L_H 与 L_O 之间的权重矩阵为 W_1、W_2，偏置矩阵为 b_1、b_2，激活函数为 f，输入层矩阵为 X，隐藏层矩阵为 H，输出值为 Y。则 X 到 H 的正向传播过程为：

$$H = f(W_1 \cdot X + b_1) \tag{11}$$

H 到 Y 的传播过程为：

$$Y = f(W_2 \cdot H + b_2) \tag{12}$$

总样本数为 N，则模型的预测误差 E 为：

$$E = \frac{1}{2}\sum_{i=1}^{N}(\vec{Y}_i - Y_i)^2 \tag{13}$$

其中 $\vec{Y}_i$ 和 Y_i 分别为第 i 个样本的实际值和预测值。

在计算出误差结果后，将 X 视为固定值，把 E 看作全体权重值 W 的函数，计算 E 的梯度 ∇E，利用梯度下降法更新权重

$$W(s+1) = W(s) - \eta \nabla E \tag{14}$$

其中 η 为学习率，s 为迭代次数。∇E 由 E 对权重矩阵中的 w 的偏导数构成，因此上式可看作对每一个 w 进行更新：

$$w_{ji}^k(s+1) = w_{ji}^k(s) - \eta \frac{\partial E}{\partial w_{ji}^k} \tag{15}$$

其中 w_{ji}^k 表示第 k 层中第 j 个神经元与第 $k+1$ 层中第 i 个神经元之间的权重。

具体来说，本模型将 I 片晶圆上的每一个点的高度 $h_{\alpha r}^i$ 视为一个样本，其中 i 表示晶圆号，α 表示待预测点的所在角度，r 表示其与中心点距离，模型构建步骤如下。

步骤一：首先对每片晶圆的数据进行校正，即：

$$h_{\alpha r_corrected}^i = h_{\alpha r}^i - h_{centre}^i \tag{16}$$

其中 h_{centre}^i 为第 i 片 wafer 中心点的高度。

步骤二：计算不同晶圆中每一位置的数据点高度的方差 σ 和每两处位置的数据点高度的皮尔逊相关系数 ρ，用于表示两点之间的相关性：

$$h_{\alpha r_mean} = \frac{\sum_{i=0}^{I} h_{\alpha r_corrected}^i}{I} \tag{17}$$

$$\sigma_{\alpha r} = \sqrt{\frac{\sum_{i=0}^{I} h_{\alpha r_corrected}^i - h_{\alpha r_mean}}{I-1}} \tag{18}$$

$$\rho_{\alpha_1 r_1, \alpha_2 r_2} = \frac{cov((\alpha_1\ r_1), (\alpha_2\ r_2))}{\sigma_{\alpha_1 r_1}\sigma_{\alpha_2 r_2}} \tag{19}$$

其中 $cov((\alpha_1\ r_1), (\alpha_2\ r_2))$ 为 $(\alpha_1\ r_1)$ 和 $(\alpha_2\ r_2)$

处高度的协方差。

方差结果表示极坐标(α,r)处的高度在所有晶圆中的不确定度如图 3,不确定性越高对模型提供的信息也就越多,因此选择方差最高处作为模型的第一个关键点。

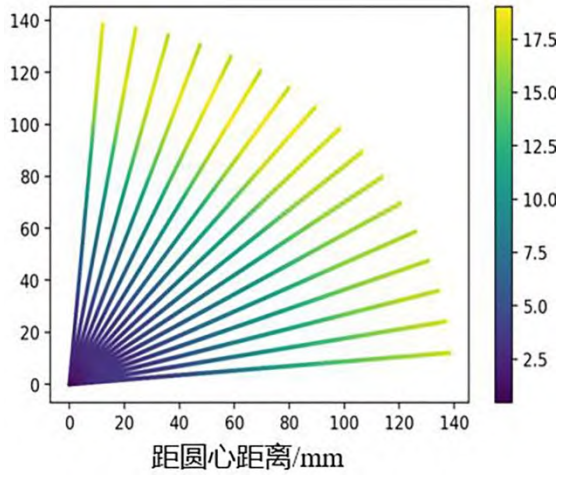


图 3　不同位置上晶圆高度方差分布图

Fig. 3 Wafer height variance distribution on different positions

步骤三:将所有样本转换为包含 n 个关键点高度特征向量 X 及对应高度 Y ,

$$X = (\alpha, r, h_1, h_2 \cdots h_n)$$

$$Y = h^i_{\alpha r_corrected}$$

并按照 0.8:0.1:0.1 的比例将所有样本晶圆划分为训练集、验证集和测试集。

步骤四:模型在训练集上学习后,在验证集上优化超参数并利用测试集对模型进行验证,得到曲面预测结果 $h^i_{\alpha r_predict}$ 。之后计算晶圆中某一位置的预测误差绝对值之和,用于表示模型对该位置的预测精度。

$$e_{\alpha r} = \sum_{i=0}^{I} \left| h^i_{\alpha r_corrected} - h^i_{\alpha r_predict} \right| \tag{20}$$

根据最终预测精度和步骤二中计算出的方差构建代理函数 $f(\alpha,r)$,用于评估点 (α,r) 的价值:

$$f(\alpha,r) = \frac{{e_{\alpha r}}^2}{\mathrm{Max}\left(\cup_{i=1}^{n} \rho((\alpha,r), keypoint_i) \right)} \tag{21}$$

将代理函数目标值最高,即与已有关键点关联性较小、预测误差较大的坐标,作为新的关键点,加入特征向量 X 的构建中,并重复步骤三、四,迭代至模型整体精确度不再提升。

经实验验证,晶圆关键点数量为 10 左右时,模型整体已经有较高的预测精度。

1.4　结果与分析

收集 250 片晶圆终抛后的厚度数据作为神经网络模型数据库进行训练集、验证集、测试集的模型评估计算,利用上述方法进行特征点选取、建模和预测后,以平均绝对误差(Mean Absolute Error,MAE)为指标评估模型精度,精度随关键点数量变化如图 4 所示:

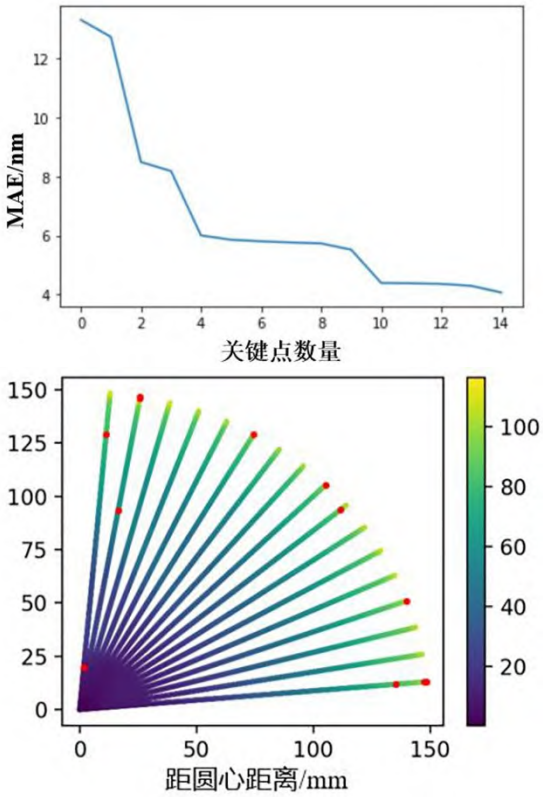


图 4　随关键点数量变化的模型评估曲线及最终关键点位置

Fig. 4 Model evaluation curve with the number of key points and final key point location

随着关键点数量的增加,模型精度不断提高,并在关键点数到达 10 个时趋于平稳。在关键点数量增至 14 时,测试集的 MAE 降至 4.07nm,已经低于当前工艺能力对平坦度的控制范围,继续增大关键点数量提升预测精度对指导工艺意义不大,并且会牺牲测量时间。因此,将 50 片测试集晶圆的 14 点厚度数据输入,通过模型模拟出第一象限曲面并将

其与实际值对比(图 5),通过直方图可以看出,整体偏差处于较低范围,集中在±20nm 区间内:

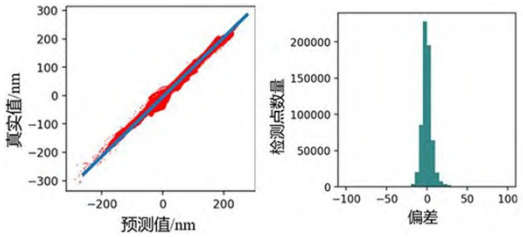


图 5 预测输出值偏差分布

Fig. 5 Predicted output value deviation distribution

通过计算每一个坐标处的 MAE,得到预测误差分布图,如图 6 所示,由于硅片形貌随特征为越靠近边缘处其不确定性越大,因此总体趋势为 MAE 随半径增大逐渐升高。然而,当位置随着半径增大到接近硅片边缘时,由于关键点分布较为密集,在模型的训练过程中相对能够获取更多的信息,因此 MAE 显著下降。

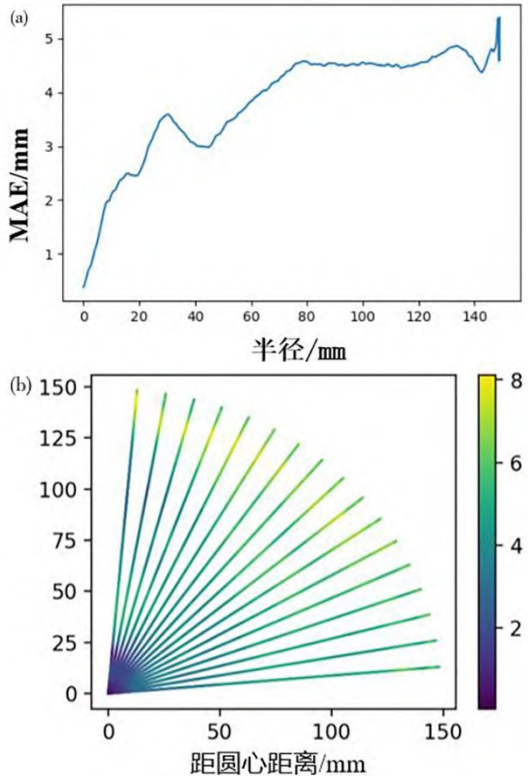
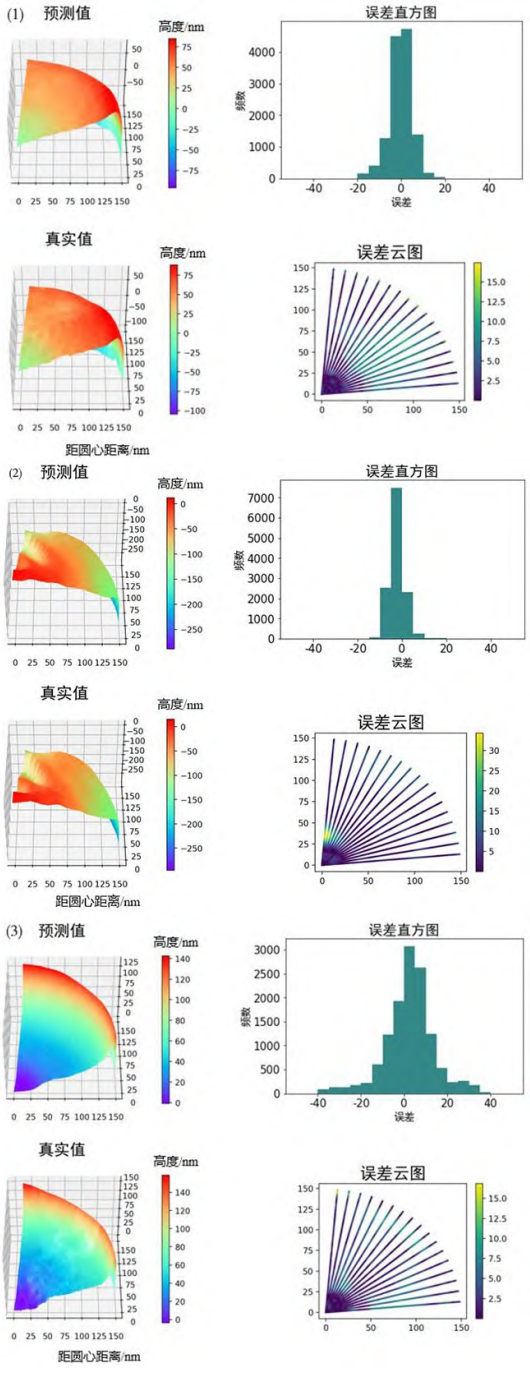


图 6 (a):MAE 随半径变化(b):MAE 随位置变化分布图

Fig. 6 (a):MAE variation with radius; (b):MAE variation distribution with position

最终根据 14 个关键点预测得到的 3D 形貌与实际形貌对比图及误差分布如图 7 所示,对测试集中 5 个样本预测形貌与实际形貌进行了对比。对比结果中可以看出模型预测的晶圆厚度分布与真实值厚度分布的 3D 形貌图像基本一致,并且其预测误差的直方图基本呈期望值 μ=0 的正态分布形状,这表明了模型预测值具有高准确性,并且预测误差的范围基本控制在±20nm 的区间内,证明了模型的有效性,能够达到较高精度的预测。



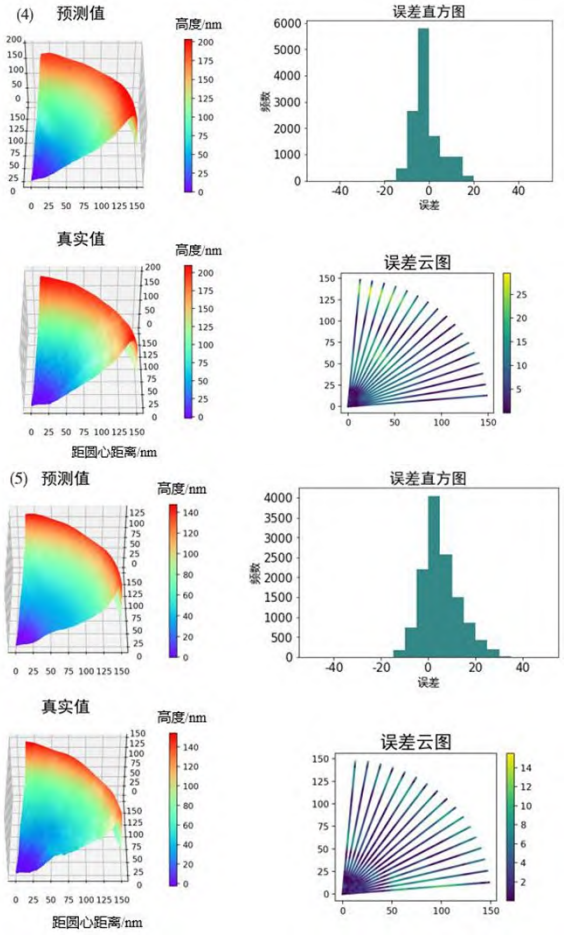


图 7　测试集中 5 个样本预测形貌与实际形貌对比

Fig. 7 Comparison of predicted shape and actual shape for 5 samples in the test set

2　结论

本研究采用神经网络算法建立了晶圆表面轮廓预测模型,提高了量测的效率同时能具有高量测精度。模型极大的减少了表征晶圆形貌所需的输入信息,以晶圆上更少的指定坐标厚度来得到高精度的预测结果,并输出晶圆 3D 形貌图像。此技术仅需 10s 左右量测 14 个关键点的厚度信息,便可得到绝对平均误差为 4.07nm 的预测值,该误差低于当前工艺能力对平坦度的调控范围,对指导工艺调控具有实际意义,并降低了量测硬件的技术要求,将来有望通过与平坦化机台的集成实现晶圆在抛光工艺过程中的形貌快速反馈。

参考文献:

[1]SCHMITZ T L, DAVIES A D, EVANS C, et al. Silicon wafer thickness variation measurements using the National Institute of Standards and Technology infrared interferometer [J]. Optical Engineering, 2003, 42(8): 2281-2290.

[2]JANSEN M J, HAITJEMA H, SCHELLEKENS P H J. A scanning wafer thickness and flatness interferometer [J]. Proceedings of SPIE - The International Society for Optical. 2004, 5252:334-345.

[3]VUKKADALA P, TURNER K T, SINHA J K. Impact of Wafer Geometry on CMP for Advanced Nodes[J]. Journal of The Electrochemical Society, 2011, 158(10): H1002.

[4]郭福东,唐锋,卢云君等. 子孔径拼接干涉的快速调整及测量[J]. 光学精密工程,2017,25(10):2682-2688.

[5]王晓茹. 基于深度学习的硅片图像缺陷检测算法研究[D]. 太原:太原理工大学, 2020.

[6]王梓铭. 基于机器视觉的晶硅片良品检测系统研究[D]. 徐州:中国矿业大学, 2021.

[7]刘媛妮, 赵国锋, 李昕. 数据虚拟化:多源异构数据集成之道[M]. 北京:人民邮电出版社有限公司, 2019.